Consumer Preferences Data Analysis in JMP: An 8-Part Case Study

Exploratory Data Analysis, Clustering, and Predictive Modeling

Table of Contents

Part 1: Salary Distribution Analysis

Objective

To explore the distribution of consumer salaries in the *Consumer Preferences* dataset and identify central tendency, variability, and the effect of outliers.

Methodology

- Used **Histogram with Outlier Box Plot** (Analyze → Distribution in JMP).
- Computed summary statistics (mean, median, quartiles, standard deviation).
- Identified and excluded outliers (extremely high salaries).
- Re-ran the histogram on the adjusted dataset to compare results.
- Examined **skewness** using boxplot shape and interquartile range (IQR).

Findings

1. **Average Salary**
    - Mean salary: **$57,983.88**
    - Indicates the typical income level in the dataset, though influenced by high outliers.
2. **Presence of Outliers**
    - A number of consumers reported salaries above **$150,000–$200,000**, significantly higher than the rest.
    - These outliers skew the distribution and inflate the mean.
3. **After Removing Outliers**
    - New mean salary: **$51,948.62**

    - Median salary: **$46,500**
    - Quartiles: Q1 = $35,000, Q3 = $65,000
    - Distribution is more representative of the majority of consumers.
4. **Interquartile Range (IQR)**
    - IQR = Q3 – Q1 = **$30,000**
    - Reflects the middle 50% of consumer salaries.
5. **Skewness**
    - The salary distribution is **positively skewed (right-skewed)**.
    - Long right tail caused by high-income outliers.
    - Median lies closer to Q1 than Q3, confirming skewness.

Business Insight

- The majority of consumers fall within the **$35K–$65K salary range**.
- High-income individuals represent a small segment but can distort averages.
- For business strategy, **median salary is a better measure** of central tendency than the mean.
- Outliers should be treated carefully in predictive modeling, as they can bias regression results and clustering.

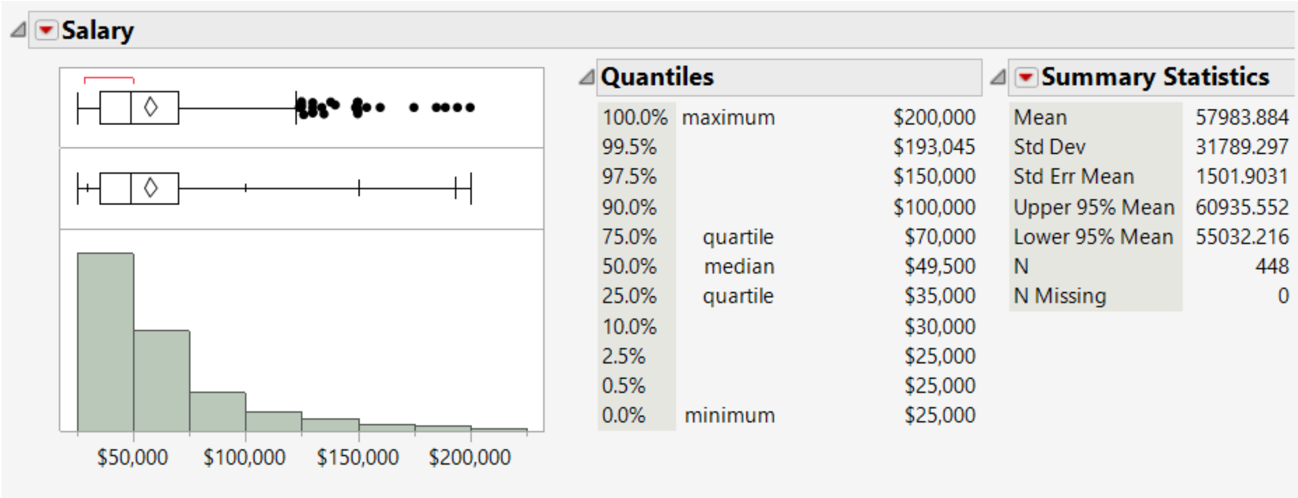
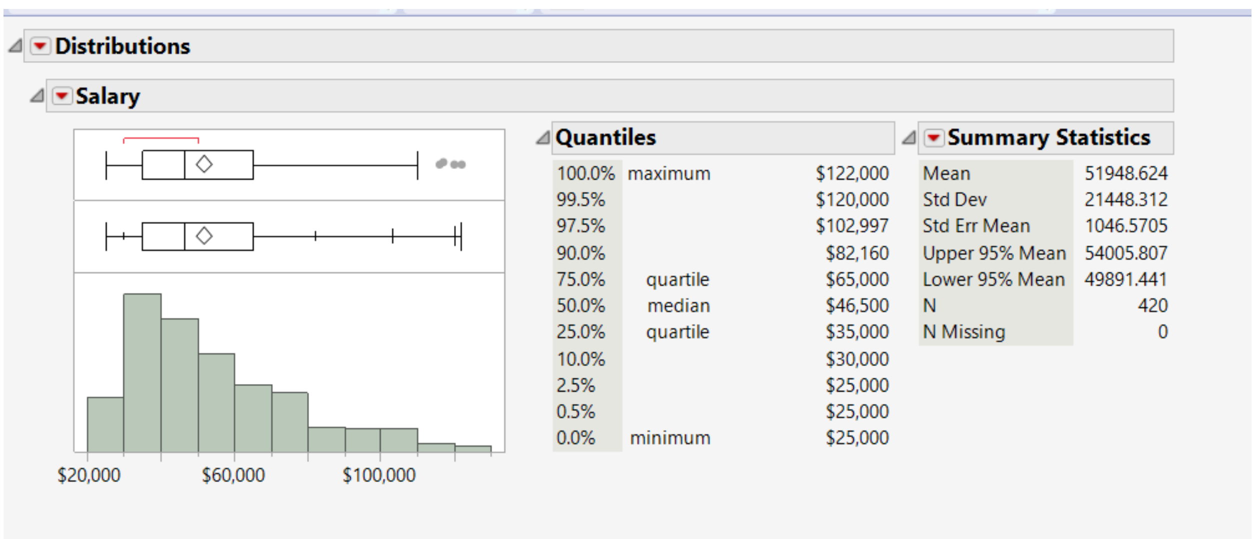


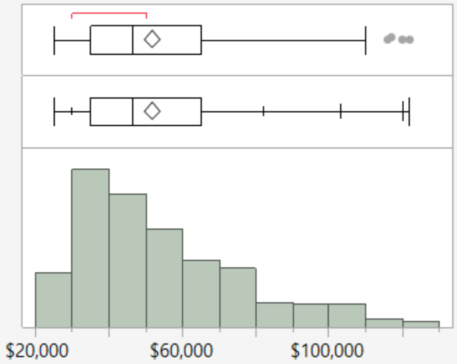
Fig: Histogram - Salary

Average salary of consumers : $57983.884. There are some consumers who have extremely high salaries compared to other consumers.



Distributions

Salary



Quantiles

100.0%	maximum	$122,000
99.5%		$120,000
97.5%		$102,997
90.0%		$82,160
75.0%	quartile	$65,000
50.0%	median	$46,500
25.0%	quartile	$35,000
10.0%		$30,000
2.5%		$25,000
0.5%		$25,000
0.0%	minimum	$25,000

Summary Statistics

Mean	51948.624
Std Dev	21448.312
Std Err Mean	1046.5705
Upper 95% Mean	54005.807
Lower 95% Mean	49891.441
N	420
N Missing	0

Fig : Histogram - Salary (outliers excluded)

We have outliers because there are 28 data points that are bigger than Q3+1.5*IQR (122,500) (-2)

IQR = Q3 - Q1

IQR = $65,000 - $35,000

IQR = $30,000

Skewed Right (positive skew): Longer right whisker, median closer to the lower quartile.

Part 2: Consumer Demographics Analysis

Objective

To analyze the gender and marital status distribution of consumers and compare visualization techniques (bar charts vs histograms). This helps understand the composition of the dataset and highlights how categorical versus continuous variables are best represented.

Methodology

- Created **Bar Charts** in JMP using *Graph* → *Graph Builder* to examine categorical variables:

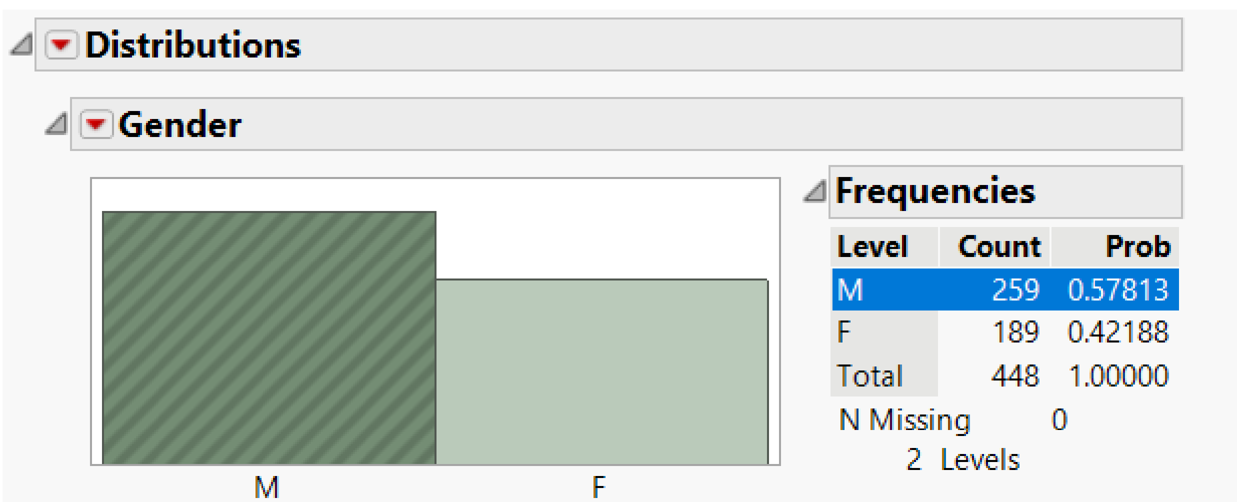
- ○ **Gender** (Male/Female)
- ○ **Marital Status** (Single/Not Single)

- Counted frequencies and calculated proportions of each category.
- Compared the use of **bar charts vs histograms** to reinforce data visualization best practices.

Findings

1. **Gender Distribution**
   - Total respondents: **448**
   - **Male:** 259 (≈ **57.8%**)
   - **Female:** 189 (≈ **42.2%**)
   - → The dataset is male-dominated, which may influence consumer behavior patterns.

2. ***Figure: Bar Chart – Gender Distribution***

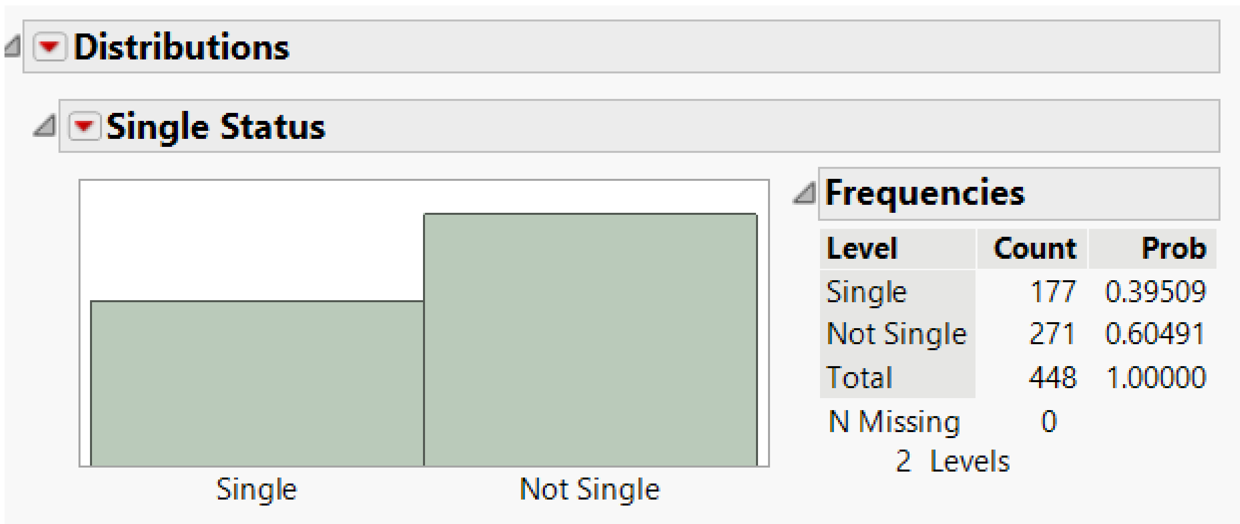


Most common gender is male - 57.813%

Male - 259, Female - 189

---

2. **Marital Status Distribution**
    - **Not Single (Married/Other):** 271 (≈ **60.5%**)
    - **Single:** 177 (≈ **39.5%**)
    - → The majority of consumers are not single, suggesting that family dynamics may play a role in product usage and purchasing decisions.
3. *Figure: Bar Chart – Marital Status Distribution*



Most common status is not single/ married - 60.491%
Single - 177, Not single - 271

---

3. **Visualization Comparison: Histogram vs Bar Chart**
    - **Data Type:**
        - Histogram: used for **continuous variables** (e.g., salary, age).
        - Bar Chart: used for **categorical variables** (e.g., gender, marital status).
    - **Bar Spacing:**
        - Histogram bars touch (continuous bins).
        - Bar Chart bars are separated (distinct categories).

- **X-axis Representation:**
  - Histogram X-axis shows **ranges/intervals**.
  - Bar Chart X-axis shows **discrete categories**.

4. → Choosing the correct visualization type ensures that the data story is accurate and intuitive for stakeholders.

Business Insight

- The dataset is skewed toward **male and married consumers**.
- Marketing strategies targeting households or family-oriented products may align with the dominant demographic.
- Visualization literacy (knowing when to use bar charts vs histograms) is essential for presenting results clearly to decision-makers.

Part 3: Relationship Between Age, Floss Cost, and Salary

Objective

To examine the relationship between consumers' age and two behavioral/financial variables—**floss cost** and **salary**—using scatterplots with fitted trend lines. Additionally, a 3D scatterplot was created to visualize all three variables simultaneously.

Methodology

- Created scatterplots in JMP (*Graph → Scatterplot Matrix*) with **fit lines** added.
- Variables analyzed:
  - Floss Cost (Y) vs. Age (X)
  - Salary (Y) vs. Age (X)

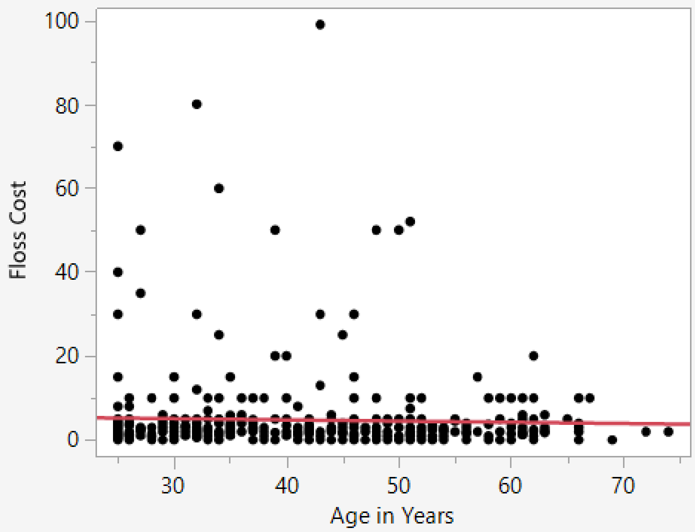
- Constructed a **3D scatterplot** (Age, Salary, Floss Cost) for multi-variable visualization.

Findings

1. **Floss Cost vs Age**
    - Scatterplot shows that floss cost is **constant across age groups**.
    - Fit line is flat, confirming **no significant relationship** between age and floss expenditures.
    - Reason: flossing products are sold at fixed prices, so cost does not vary by demographic.

2. *Figure: Scatterplot – Floss Cost vs Age with Fit Line*

Bivariate Fit of Floss Cost By Age in Years



Linear Fit

Linear Fit

Floss Cost = 5.9559938 - 0.0285353*Age in Years

Summary of Fit

RSquare	0.001268
RSquare Adj	-0.00097
Root Mean Square Error	9.908148
Mean of Response	4.800313
Observations (or Sum Wgts)	448

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	55.572	55.5720	0.5661
Error	446	43784.441	98.1714	**Prob > F**
C. Total	447	43840.013		0.4522

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.9559938	1.605787	3.71	0.0002*
Age in Years	-0.028535	0.037927	-0.75	0.4522

---

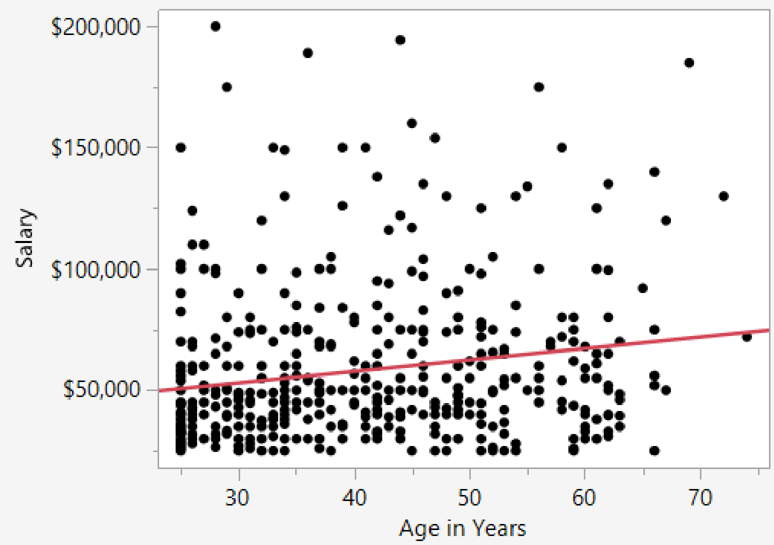
2. **Salary vs Age**

    - Scatterplot indicates a **positive relationship**: as age increases, salary tends to rise.

- The fit line slopes upward, though with considerable spread, meaning age explains some but not all variation in salary.
- Younger consumers cluster in lower salary ranges, while older consumers are more likely to appear in higher salary brackets.

3. *Figure: Scatterplot – Salary vs Age with Fit Line*

Bivariate Fit of Salary By Age in Years



Linear Fit

Linear Fit

Salary = 38842.123 + 472.63608*Age in Years

Summary of Fit

RSquare	0.03375
RSquare Adj	0.031584
Root Mean Square Error	31283.26
Mean of Response	57983.88
Observations (or Sum Wgts)	448

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.5246e+10	1.525e+10	15.5783
Error	446	4.3647e+11	978642318	**Prob > F**
C. Total	447	4.5172e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38842.123	5069.994	7.66	<.0001*
Age in Years	472.63608	119.7477	3.95	<.0001*

Business Insight

- **Salary grows with age**, which may indicate increased purchasing power among older consumers.
- **Floss cost is independent of age**, suggesting oral hygiene purchases are consistent and non-discretionary across demographics.
- Companies can segment consumers by **salary/age brackets** for targeted marketing of premium vs. value-added products, but floss pricing strategies should remain standardized.

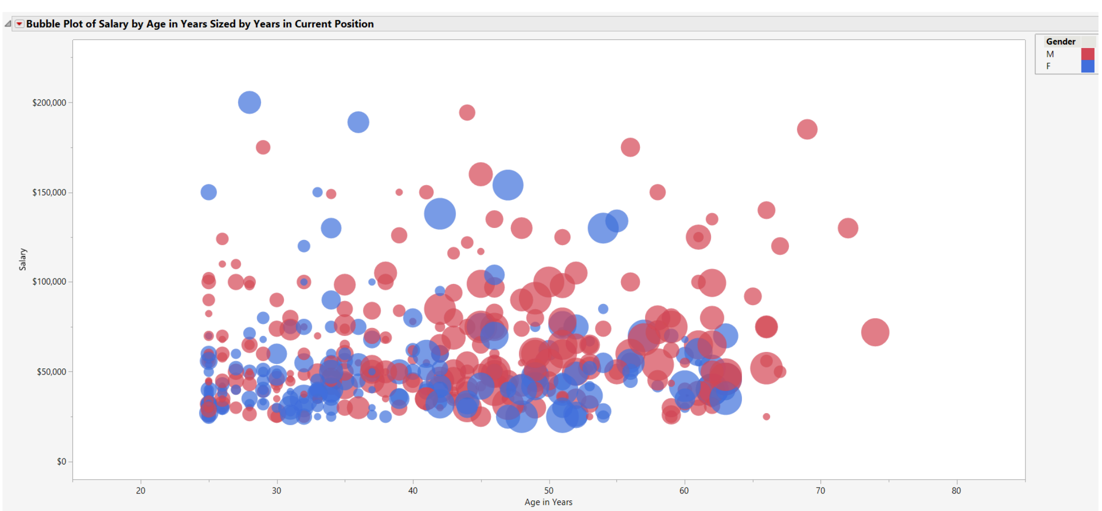
Part 4: Multivariate Analysis Using Bubble Plot

Objective

To explore the combined relationship between **salary, age, and years in current position**, with **gender** included as a categorical dimension. The goal is to understand how multiple factors interact simultaneously and identify meaningful consumer subgroups.

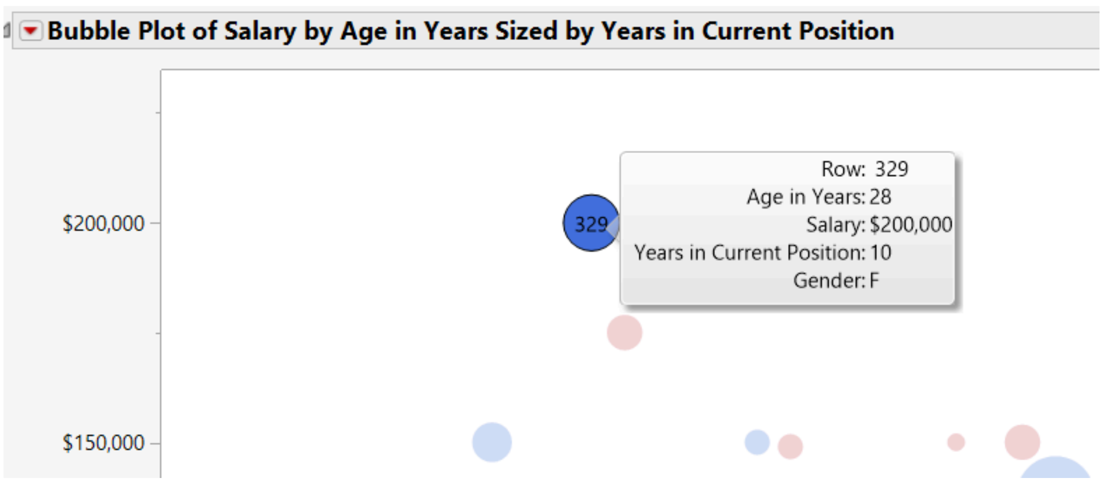
Methodology

- Constructed a **Bubble Plot** in JMP (*Graph → Bubble Plot*).



- Variables used:
    - **X-axis:** Age
    - **Y-axis:** Salary
    - **Bubble Size:** Years in Current Position
    - **Color:** Gender

Example:



- Animated and static plots were examined to interpret demographic and career dynamics.

Findings

1. **Salary vs Age**
    - Clear upward trend: **older consumers generally earn higher salaries**.
    - Younger individuals cluster in the lower-salary range.
2. **Tenure (Bubble Size)**
    - Larger bubbles (longer years in position) are concentrated in **older age groups**, confirming a link between career stability and salary growth.

    - Smaller bubbles (shorter tenure) appear in younger respondents.

3. **Gender Dynamics (Bubble Color)**
    - Both genders are represented across all income levels.
    - The visual did not reveal strong gender-specific differences, though male representation is slightly higher overall (consistent with Part 2 findings).

4. **Overall Multivariate Pattern**
    - The bubble plot highlights **career progression**: as individuals age, their **years in current position** and **salary** both increase.
    - Younger employees are early in their careers (small bubbles, lower salaries), while older employees demonstrate longer tenure and higher income.

Business Insight

- **The career stage is a strong driver of salary growth.** Marketing campaigns for higher-value or premium products may be better targeted to older, longer-tenured professionals.
- Younger, lower-salary consumers may represent a **price-sensitive segment** requiring affordable offerings.
- **Gender segmentation is less critical** in salary-based targeting but may still play a role in product messaging.

Part 5: Gender Differences in Toothpaste Spending

Objective

To investigate whether there are differences in toothpaste expenditure between male and female consumers, using multiple visualization techniques for comparison.

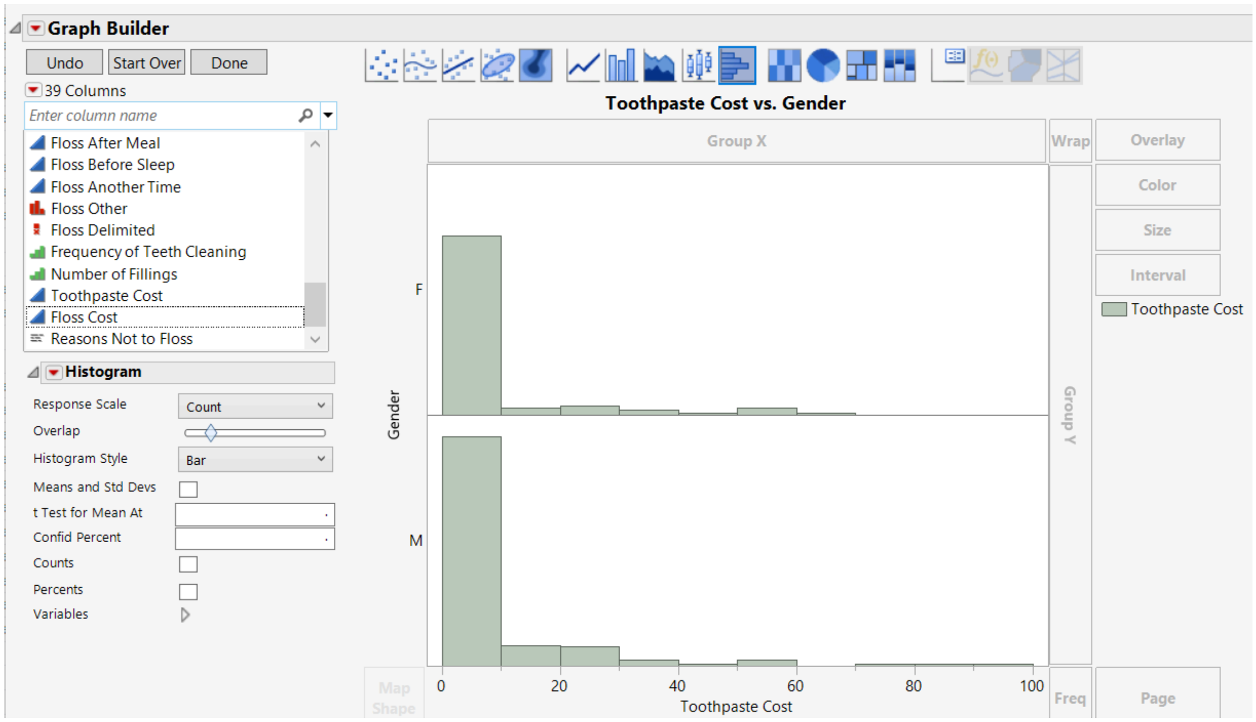
Methodology

- Used **JMP Graph Builder** (*Graph → Graph Builder*).
- Constructed two visualizations:
  1. **Bar Chart**: Toothpaste cost (X-axis) vs Gender (Y-axis).
  2. **Pie Chart**: Proportional share of toothpaste spending by gender.
- Compare results from both visualizations to assess consistency.

Findings

1. **Bar Chart**
   - Male consumers show **higher frequency** in the upper cost ranges compared to females.
   - Female consumers are more concentrated in lower cost brackets.

2. *Figure: Bar Chart – Toothpaste Cost vs Gender*

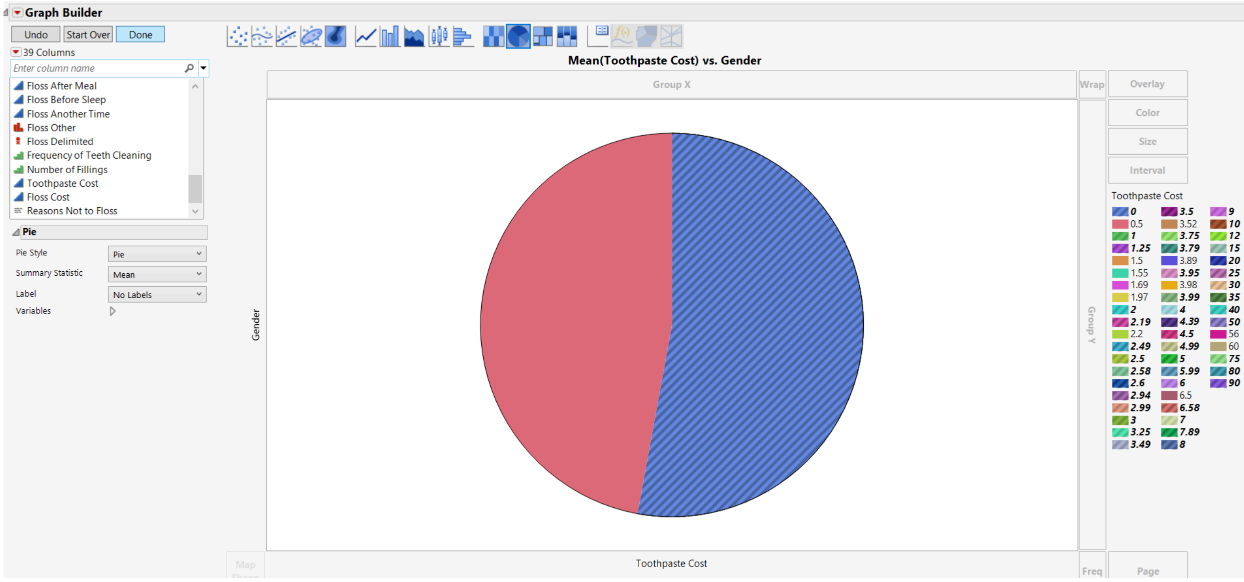


---

2. **Pie Chart**

   - Confirms the bar chart result:
   - **Male consumers spend more overall** on toothpaste than female consumers.
   - While both genders contribute significantly, the male share is slightly dominant.

3. *Figure: Pie Chart – Toothpaste Cost vs Gender*



Business Insight

- Toothpaste spending differs by gender, with **males spending more than females**.
- This suggests opportunities for **gender-targeted marketing**:
    - Premium or bulk toothpaste offerings may appeal more to male consumers.
    - Female consumers may be more price-conscious, making them a stronger target for **value or promotional packs**.
- By validating results through two different graphs, the analysis demonstrates consistency and strengthens confidence in the insight.

Part 6: Customer Segmentation Using K-Means Clustering

Objective

To segment consumers into groups based on **age, years in current position, and years with current employer** using the **K-Means clustering method**. This analysis helps identify patterns of career progression and job stability across different consumer profiles.

Methodology

- Applied **K-Means Clustering** in JMP (*Analyze → Clustering*).
- Number of clusters set to **3**, based on optimal CCC (Cubic Clustering Criterion).
- Variables included:
  - Age (in years)
  - Years in current position
  - Years with current employer
- Results visualized using both **Biplot** and **Biplot 3D** for interpretability.

Findings

1. **Cluster Distribution**
   - **Cluster 1:** 71 individuals (older, long tenure).
   - **Cluster 2:** 220 individuals (younger, very short tenure).
   - **Cluster 3:** 157 individuals (mid-career, moderate tenure).

2. *Table: Cluster Means & Standard Deviations*

Iterative Clustering

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	3	1.44879	Optimal CCC

Columns Scaled Individually

Control Panel

Method: K Means Cluster

Number of Clusters: 3

Range of Clusters (Optional): .

Go

☐ Single Step

☐ Use within-cluster std deviations

☐ Shift distances using sampling rates

K Means NCluster=3

Columns Scaled Individually

Cluster Summary

Cluster	Count
1	71
2	220
3	157

Step	Criterion
15	0

Cluster Means

Cluster	Age in Years	Years in Current Position	Years at Current Employer
1	54.056338	14.7746479	24.5070423
2	30.2318182	3.13636364	4.08181818
3	48.7579618	6.93630573	8.80254777

Cluster Standard Deviations

Cluster	Age in Years	Years in Current Position	Years at Current Employer
1	7.30403578	4.84171191	6.96646543
2	5.02320853	2.03816884	2.7489893
3	8.72758368	3.6497649	5.00501594

---

2. **Cluster Profiles**

- **Cluster 1 – Experienced and Stable Professionals**
    - Avg Age: ~54 years

- Avg Years in Position: ~14.8
- Avg Years with Employer: ~24.5
- Interpretation: This cluster represents **older, highly experienced employees** with long-term stability in both role and employer.

- **Cluster 2 – Early-Career Employees**
  - Avg Age: ~30 years
  - Avg Years in Position: ~3.1
  - Avg Years with Employer: ~4.0
  - Interpretation: **Young employees**, likely in early career stages, with shorter tenure and higher mobility.
- **Cluster 3 – Mid-Career Professionals**
  - Avg Age: ~49 years
  - Avg Years in Position: ~6.9
  - Avg Years with Employer: ~8.8
  - Interpretation: **Mid-career employees** balancing moderate tenure with career advancement.
- *Figure: Biplot – Clustering Results (2D)*

Iterative Clustering

K Means NCluster=3

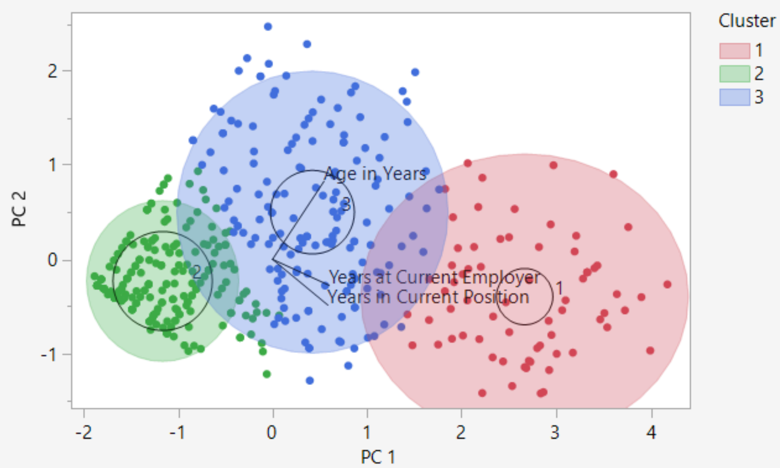
Cluster Means

Cluster	Age in Years	Years in Current Position	Years at Current Employer
1	54.056338	14.7746479	24.5070423
2	30.2318182	3.13636364	4.08181818
3	48.7579618	6.93630573	8.80254777

Cluster Standard Deviations

Cluster	Age in Years	Years in Current Position	Years at Current Employer
1	7.30403578	4.84171191	6.96646543
2	5.02320853	2.03816884	2.7489893
3	8.72758368	3.6497649	5.00501594

Biplot

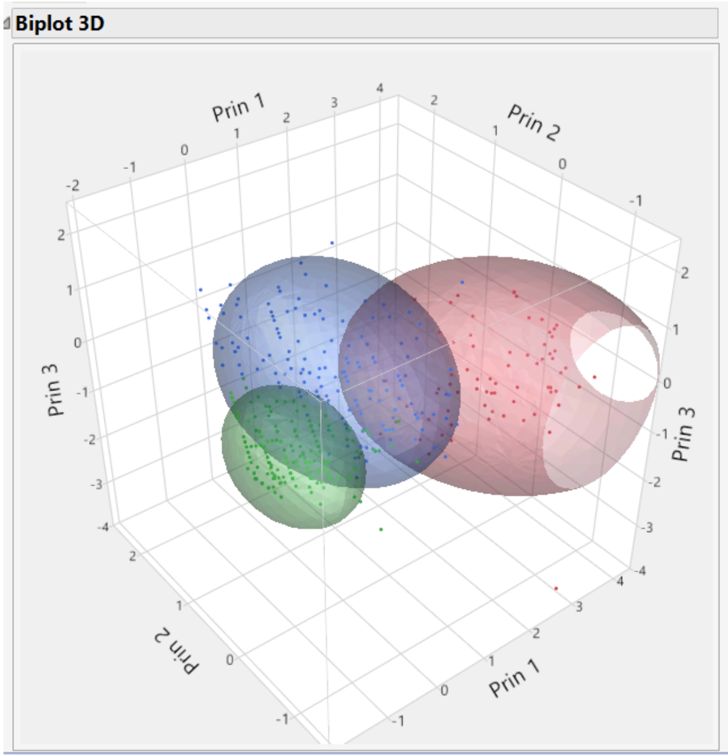


Select principal components PC 1 PC 2

Save Colors to Table

Eigenvalues
2.1456459 0.5225865 0.3317675

- *Figure: Biplot 3D – Clustering Results (3D visualization)*



Insights

- **Experience-Tenure Correlation**
    - Older employees (Cluster 1) are strongly associated with job stability, remaining with the same employer and position for decades.
    - This group may represent **loyal, stable consumers** with predictable spending habits.
- **Younger Employees Show Higher Mobility**
    - Cluster 2 reveals that younger employees tend to switch positions and employers frequently.
    - Reflects career exploration, instability, or higher risk tolerance. Businesses targeting this group should expect **dynamic preferences and less brand loyalty**.
- **Mid-Career Professionals are a Transitional Segment**

- Cluster 3 balances stability and growth, making them a **prime target for mid-range or aspirational products**.

Business Insight

Clustering allows businesses to segment their workforce-related consumer base into **career-stage profiles**.

- **Cluster 1:** Target with stability-oriented, premium products.
- **Cluster 2:** Focus on entry-level, affordable, or flexible offerings.
- **Cluster 3:** Position aspirational products for professionals transitioning into higher salary brackets.

Part 7: Predictive Modeling of Tooth Brushing Behavior

Objective

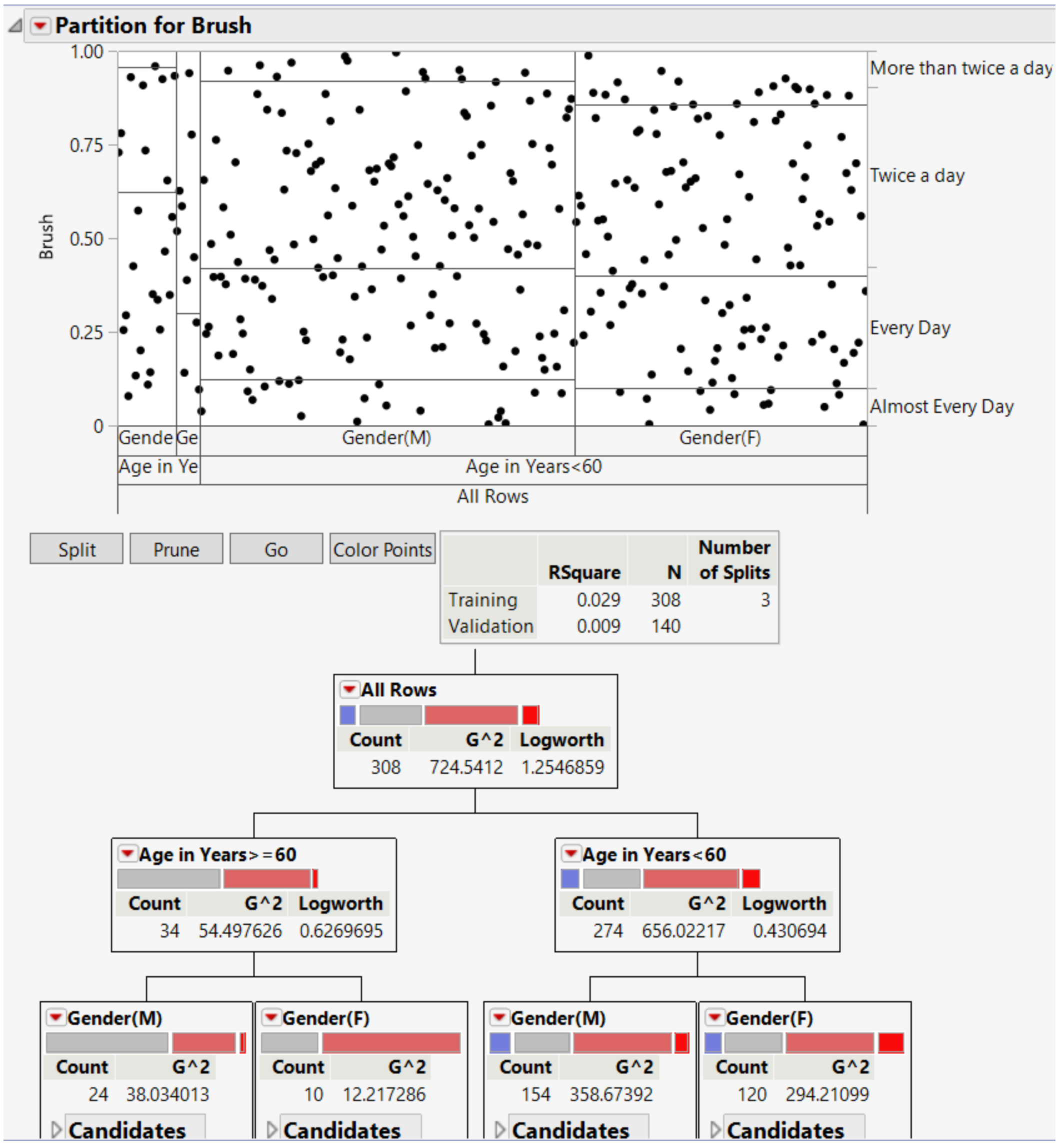
To analyze **tooth brushing frequency** across consumers using a **decision tree partition model**. The goal was to classify behavior based on **age** and **gender**, and to identify key demographic patterns in oral hygiene habits.

Methodology

- Used **Partitioning** in JMP (*Analyze → Predictive Modeling → Partition*).
- Target variable: **Brushing Frequency** (`Brush`).
- Predictor variables: **Age in Years** and **Gender`.
- Parameters:
    - Validation set: 30% of data.
    - Number of splits: 3.

- Visualization: **Decision Tree with Split Probabilities** to interpret predicted brushing frequencies by subgroup.

👉 **Figure 1**



Findings

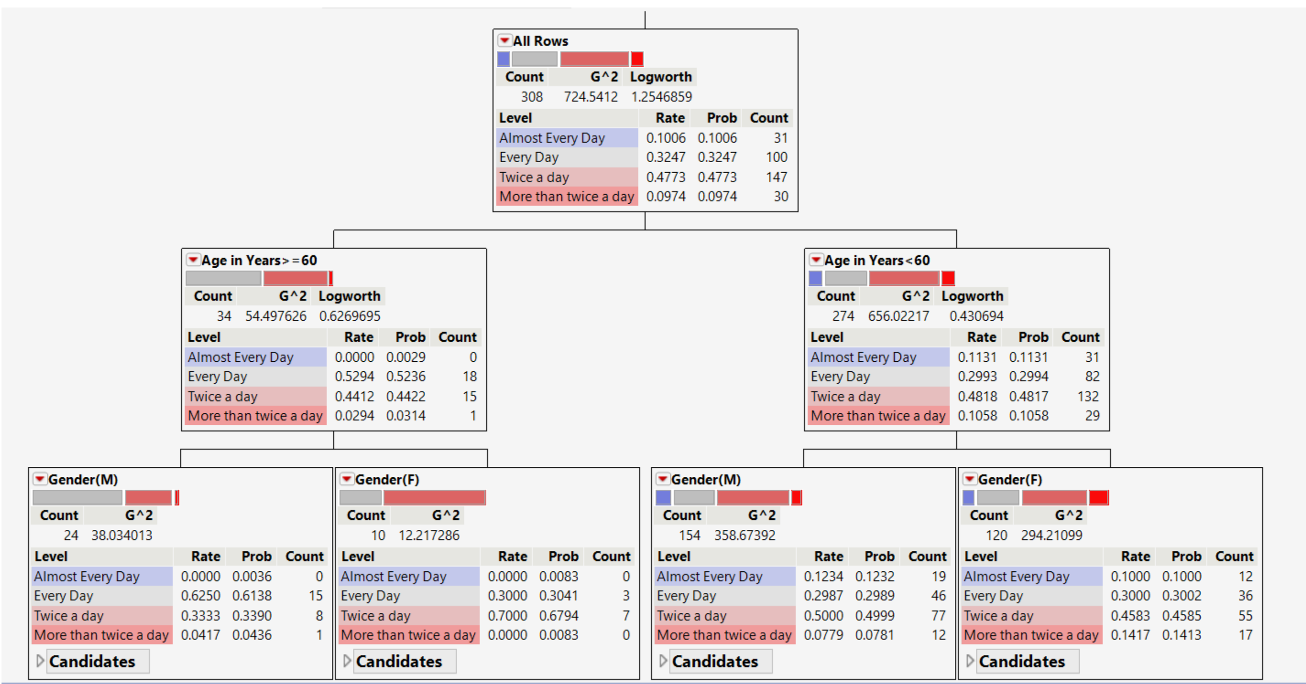
1. Age Factor

- **Older Consumers (≥ 60 years)**

- Brush less frequently overall.
- Most brush **Every Day (52.94%)** or **Twice a Day (44.12%)**.
- None reported “More than Twice a Day.”

- **Younger Consumers (< 60 years)**
  - More variation in brushing habits.
  - Significant share brush **Twice a Day (48.18%)**, followed by **Every Day (29.93%)**.
  - Slightly higher tendency toward increased brushing frequency than older groups.

👉 **Figure 2 here:** *Split Probability View (with proportions/probabilities for each node)*



---

2. Gender Factor

- **Males ≥ 60**
  - 62.5% brush **Every Day**, 33.3% brush **Twice a Day**.

- **Females ≥ 60**
    - 50% brush **Twice a Day**, 30% brush **Every Day**.
- **Males < 60**
    - 45.83% brush **Twice a Day**, 29.87% brush **Every Day**.
    - 7.79% brush more than twice daily.
- **Females < 60**
    - 45.85% brush **Twice a Day**, 30.02% brush **Every Day**.
    - A small group brushes more than twice daily.

Key Insights

- **Age Impact**: Younger consumers brush more frequently than older ones, especially twice daily.
- **Gender Impact**: Patterns are similar between males and females, though **females < 60 slightly exceed males** in “Twice a Day” brushing.
- **Practical Implications**:
    - **Younger consumers**: Emphasize lifestyle-driven oral care (whitening, freshness).
    - **Older consumers**: Focus on preventive messaging (gum health, consistency).

Part 8: Regression Analysis — Salary by Age

Objective

To explore whether a consumer’s age has a measurable relationship with their salary, and to estimate salary differences between younger and older individuals.

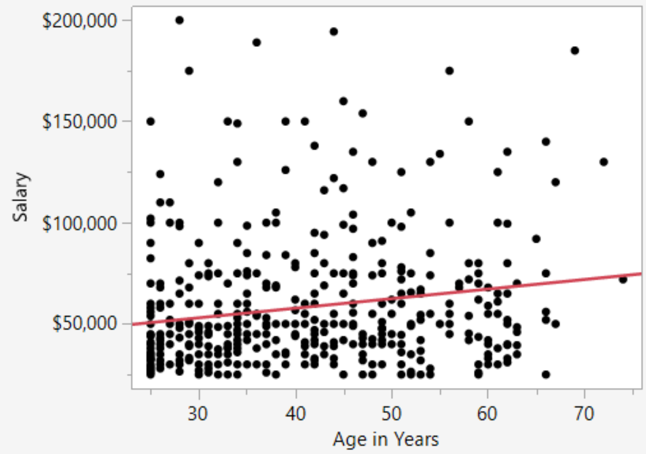
Methodology

- Applied **Simple Linear Regression** in JMP (*Analyze → Fit Y by X*).
- Response Variable: **Salary**
- Predictor Variable: **Age in Years**
- Added **Fit Line** to visualize the relationship.

Results

📌 **Figure:** *“Bivariate Fit of Salary by Age in Years”*

Bivariate Fit of Salary By Age in Years



Linear Fit

Linear Fit

Salary = 38842.123 + 472.63608*Age in Years

Summary of Fit

RSquare	0.03375
RSquare Adj	0.031584
Root Mean Square Error	31283.26
Mean of Response	57983.88
Observations (or Sum Wgts)	448

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1.5246e+10	1.525e+10	15.5783
Error	446	4.3647e+11	978642318	**Prob > F**
C. Total	447	4.5172e+11		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	38842.123	5069.994	7.66	<.0001*
Age in Years	472.63608	119.7477	3.95	<.0001*

1. **Average Salary**
   - The average salary across the sample is **$57,983.88** (from the "Mean of Response" section).
2. **Salary Difference by Age (60 vs. 25 years old)**
   - Regression equation:

$$\text{Salary} = 38842.123 + (472.63608 \times \text{Age})$$

- Salary at 60 years old: **$67,182.29**
- Salary at 25 years old: **$50,657.03**
- **Difference:** On average, a 60-year-old earns **$16,525.26 more** than a 25-year-old.

3. **Model Fit Quality**
    - The regression fit is **weak** ($R^2 = 0.03375$).
    - This indicates that **only ~3.4% of the variation in salary is explained by age**.
    - Other variables (e.g., education, industry, experience, gender, region) likely play a much stronger role.

Insights for Stakeholders

- While age shows a **positive correlation** with salary, the relationship is weak.
- The data suggests that salary determination is **multifactorial**. Age alone is not a reliable predictor, and broader labor market or demographic factors likely dominate.
- Demonstrates the ability to apply **regression modeling**, interpret coefficients, and critically evaluate model fit.

Overall Conclusion & Key Takeaways

This 8-part analysis of the *Consumer Preferences* dataset demonstrates the application of exploratory data analysis, clustering, regression, and predictive modeling techniques using JMP.

Key cross-cutting insights include:

- **Demographics matter**: The dataset skews male and not single; age strongly correlates with salary but weakly predicts behavior.
- **Segmentation is powerful**: K-Means clustering revealed three distinct career-stage groups (young/short-tenure, mid-career, and older/stable employees), each with different consumer profiles.

- **Behavioral trends**: Brushing frequency varies more by age than by gender; toothpaste spending patterns also show gender-based differences.

- **Visualization matters**: Using the right chart type (histogram vs bar chart, scatterplot vs bubble plot) ensures accurate communication to stakeholders.

- **Predictive modeling adds value**: While age alone is a weak predictor of salary, decision trees highlighted nuanced brushing behavior patterns, showing the value of applied machine learning.

Practical Implications

- Companies can tailor marketing strategies by **career stage, salary bracket, and brushing behavior**.

- **Premium products** may target stable, older professionals, while **value products** may resonate more with younger, mobile consumers.

- Insights like these bridge raw data and actionable business decisions — a crucial skill in modern data analysis.