**AMCAT EDA**

**Data Overview:**

* **Number of Rows:** 3998
* **Number of Columns:** 38
* **Data Size:** 151,924 entries
* **Missing Values:** None (all columns have 3998 non-null entries)
* **Data Types:** Integers, Floats, Datetime, and Objects (categorical)

**Uni variate analysis:**

**Categorical Columns:**

**Designation:**

* **Nunique values:** 419
* **Most frequent value:** Software Engineer (13.48%)
* **Top 3 categories:** Software Engineer (539), Software Developer (265), System Engineer (205)
* **Bottom 3 categories:** CAD Drafter (1), NOC Engineer (1), Junior Software Developer (1)

**JobCity:**

* **Nunique values:** 339
* **Most frequent value:** Bangalore (15.68%)
* **Top 3 categories:** Bangalore (627), Noida (368), Hyderabad (335)
* **Bottom 3 categories:** Tirunelvelli (1), Ernakulam (1), Nanded (1)

**Gender:**

* **Nunique values:** 2 (Male, Female)
* **Most frequent value:** Male (76.06%)
* **Top category:** Male (3041)
* **Bottom category:** Female (957)

**Degree:**

* **Nunique values:** 4
* **Most frequent value:** B.Tech/B.E. (92.55%)
* **Top 3 categories:** B.Tech/B.E. (3700), MCA (243), M.Tech./M.E. (53)
* **Bottom category:** M.Sc. (Tech.) (2)

**Specialization:**

* **Nunique values:** 46
* **Most frequent value:** Electronics and Communication Engineering (22.01%)
* **Top 3 categories:** Electronics and Communication Engineering (880), Computer Science & Engineering (744), Information Technology (660)
* **Bottom 3 categories:** Embedded Systems Technology, Polymer Technology, Power Systems and Automation (1 each)

**CollegeTier:**

* **Nunique values:** 2
* **Most frequent value:** Tier 2 (92.57%)
* **Top category:** Tier 2 (3701)
* **Bottom category:** Tier 1 (297)

**CollegeState:**

* **Nunique values:** 26
* **Most frequent value:** Uttar Pradesh (22.89%)
* **Top 3 categories:** Uttar Pradesh (915), Karnataka (370), Tamil Nadu (367)
* **Bottom 3 categories:** Goa (1), Meghalaya (2), Sikkim (3)

**CollegeCityTier:**

* **Nunique values:** 2
* **Most frequent value:** Tier 0 (69.96%)
* **Top category:** Tier 0 (2797)
* **Bottom category:** Tier 1 (1201)

**12board:**

* **Nunique values:** 13
* **Most frequent value:** CBSE (57.63%)
* **Top 3 categories:** CBSE (2304), State Board (1342), ISC (245)
* **Bottom 3 categories:** International Board (1), NIOS (1), Unknown (2)

**Numerical Columns:**

**Salary:**

* **Nunique values:** 177
* **Distribution:** Highly right-skewed (Skewness: 6.45)
* **Min:** 35,000, **Max:** 4,000,000, **Mean:** 307,699
* Outliers will be detected and removed using an interquartile range (IQR) method, removing values outside **1.5 times the IQR**.

**10percentage:**

* **Nunique values:** 851
* **Distribution:** Slightly left-skewed (Skewness: -0.59)
* **Min**: 43.0, **Max**: 97.76, **Mean**: 77.92

**GraduationYear:**

* **Nunique values:** 11
* **Distribution:** Highly left-skewed (Skewness: -63.07)
* **Min**: 0 (incorrect entries), **Max**: 2017, **Mean**: 2012.10

**English:**

* **Nunique values:** 111
* **Distribution:** Nearly normal distribution (Skewness: 0.09)
* **Min**: 180, **Max**: 875, **Mean**: 501.65

**12percentage:**

* **Nunique values:** 654
* **Distribution:** Slightly right-skewed (Skewness: 0.18)
* **Min**: 41.0, **Max**: 99.60, **Mean**: 79.24

**CollegeGPA:**

* **Nunique values:** 191
* **Distribution:** Right-skewed (Skewness: 0.56)
* **Min**: 4.5, **Max**: 10.0, **Mean**: 7.05

**Logical:**

* **Nunique values:** 92
* **Distribution:** Nearly normal distribution (Skewness: 0.08)
* **Min**: 180, **Max**: 900, **Mean**: 499.16

**Quant:**

* **Nunique values:** 92
* **Distribution:** Normal distribution (Skewness: 0.04)
* **Min**: 180, **Max**: 900, **Mean**: 499.20

**Domain:**

* **Nunique values:** 177
* **Distribution:** Slightly right-skewed (Skewness: 0.43)
* **Min**: 100.0, **Max**: 999.0, **Mean**: 536.87

**ComputerProgramming:**

* **Nunique values:** 67
* **Distribution:** Right-skewed (Skewness: 0.56)
* **Min**: 180, **Max**: 900, **Mean**: 498.45

**ElectronicsAndSemicon:**

* **Nunique values:** 59
* **Distribution:** Left-skewed (Skewness: -0.63)
* **Min**: 180, **Max**: 900, **Mean**: 492.11

**ComputerScience:**

* **Nunique values:** 72
* **Distribution:** Slightly right-skewed (Skewness: 0.32)
* **Min**: 180, **Max**: 900, **Mean**: 497.46

**MechanicalEngg:**

* **Nunique values:** 53
* **Distribution:** Left-skewed (Skewness: -0.91)
* **Min**: 180, **Max**: 900, **Mean**: 479.60

**ElectricalEngg:**

* **Nunique values:** 52
* **Distribution:** Slightly right-skewed (Skewness: 0.19)
* **Min**: 180, **Max**: 900, **Mean**: 500.24

**TelecomEngg:**

* **Nunique values:** 44
* **Distribution:** Left-skewed (Skewness: -0.72)
* **Min**: 180, **Max**: 900, **Mean**: 486.74

**CivilEngg:**

* **Nunique values:** 35
* **Distribution:** Left-skewed (Skewness: -0.86)
* **Min**: 180, **Max**: 900, **Mean**: 463.92

**Conscientiousness:**

* **Nunique values:** 153
* **Distribution:** Right-skewed (Skewness: 0.63)
* **Min**: 0.0, **Max**: 1.0, **Mean**: 0.73

**Agreeableness:**

* **Nunique values:** 136
* **Distribution:** Right-skewed (Skewness: 0.57)
* **Min**: 0.0, **Max**: 1.0, **Mean**: 0.67

**Extraversion:**

* **Nunique values:** 140
* **Distribution:** Slightly right-skewed (Skewness: 0.41)
* **Min**: 0.0, **Max**: 1.0, **Mean**: 0.58

**Bi variate and multi variate analysis**

**Analysis of the "Salary of the employees year-wise" Plot**

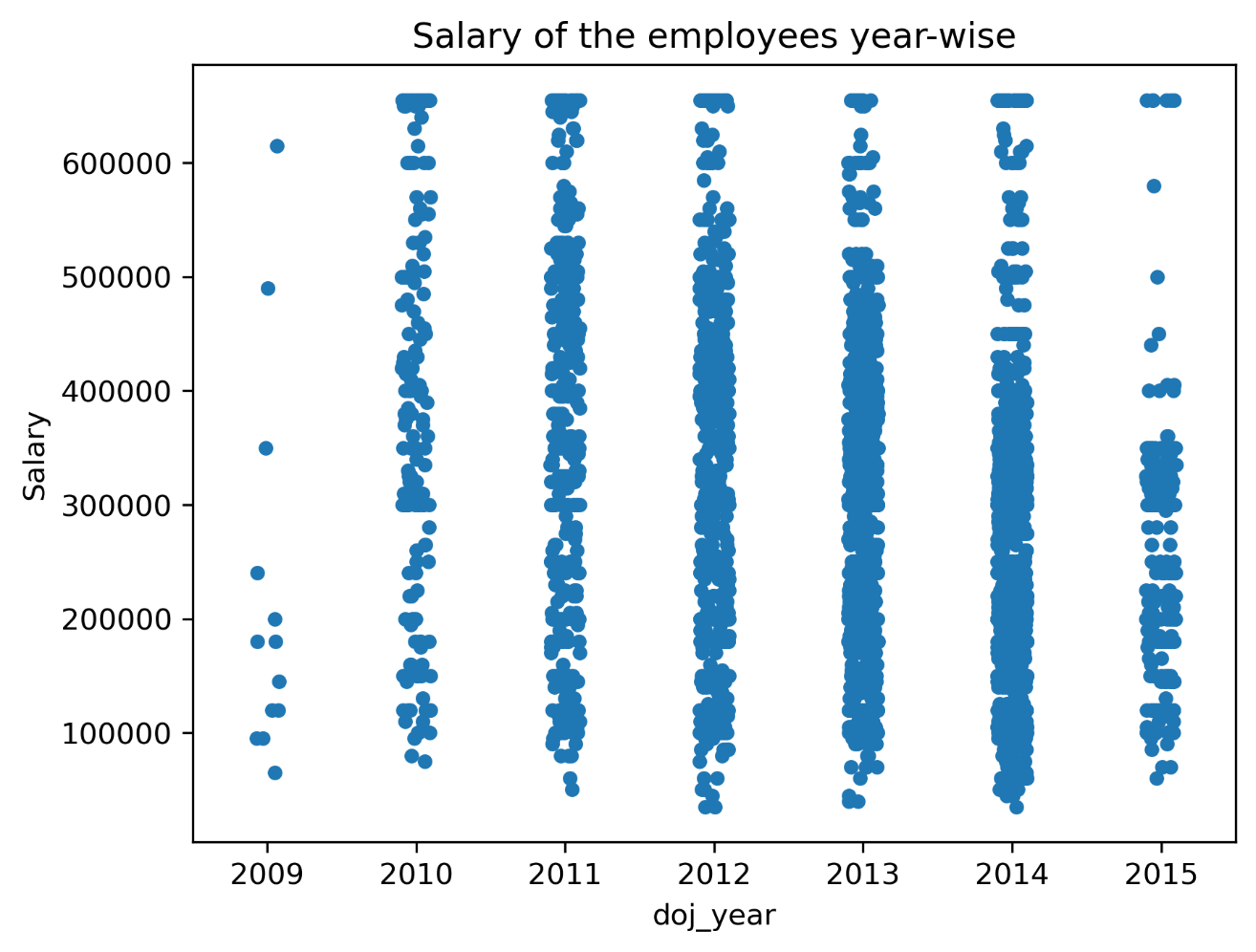
**Plot Type:** Scatter Plot

**Variables:**

* **X-axis:** DOJ Year (Date of Joining Year)
* **Y-axis:** Salary

**Observations:**

1. **Salary Distribution:** The plot shows a wide range of salaries among employees, with a significant number of employees earning between 300,000 and 400,000.
2. **Year-wise Comparison:** There appears to be a general trend of increasing salaries over the years, with the number of employees earning higher salaries increasing from 2009 to 2015. However, there is also a significant overlap in salary ranges across different years, suggesting individual variations in salary growth.
3. **Outliers:** A few outliers can be observed, particularly in the lower salary ranges, which might indicate factors such as entry-level positions, part-time roles, or regional differences in salary levels.

****

**Analysis of "12graduation year vs salary" Plot**

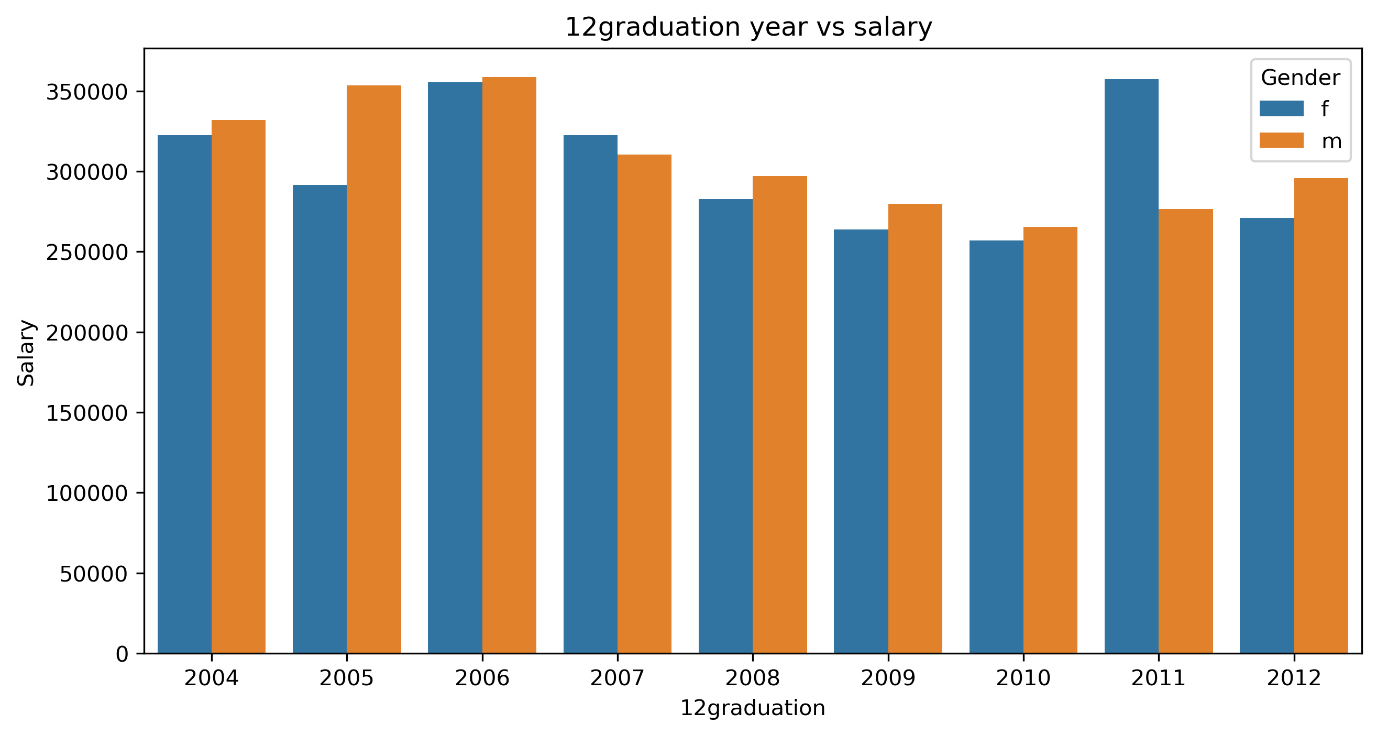
**Plot Type:** Grouped Bar Chart

**Variables:**

* **X-axis:** 12graduation (Graduation Year)
* **Y-axis:** Salary
* **Grouping Variable:** Gender (Female, Male)

**Observations:**

1. **Gender-based Salary Comparison:** The chart shows a clear comparison of salaries between male and female graduates across different graduation years.
2. **Overall Salary Trends:** There appears to be a general trend of increasing salaries over the years, with the average salary for both genders increasing from 2004 to 2012.
3. **Gender Pay Gap:** While the overall trend is similar for both genders, there seems to be a consistent gender pay gap, with male graduates generally earning higher salaries than female graduates in most years.
4. **Year-to-Year Variations:** The size of the gender pay gap varies slightly from year to year, indicating fluctuations in the degree of salary disparity.



**Analysis of "salary vs 12percentage" Plot**

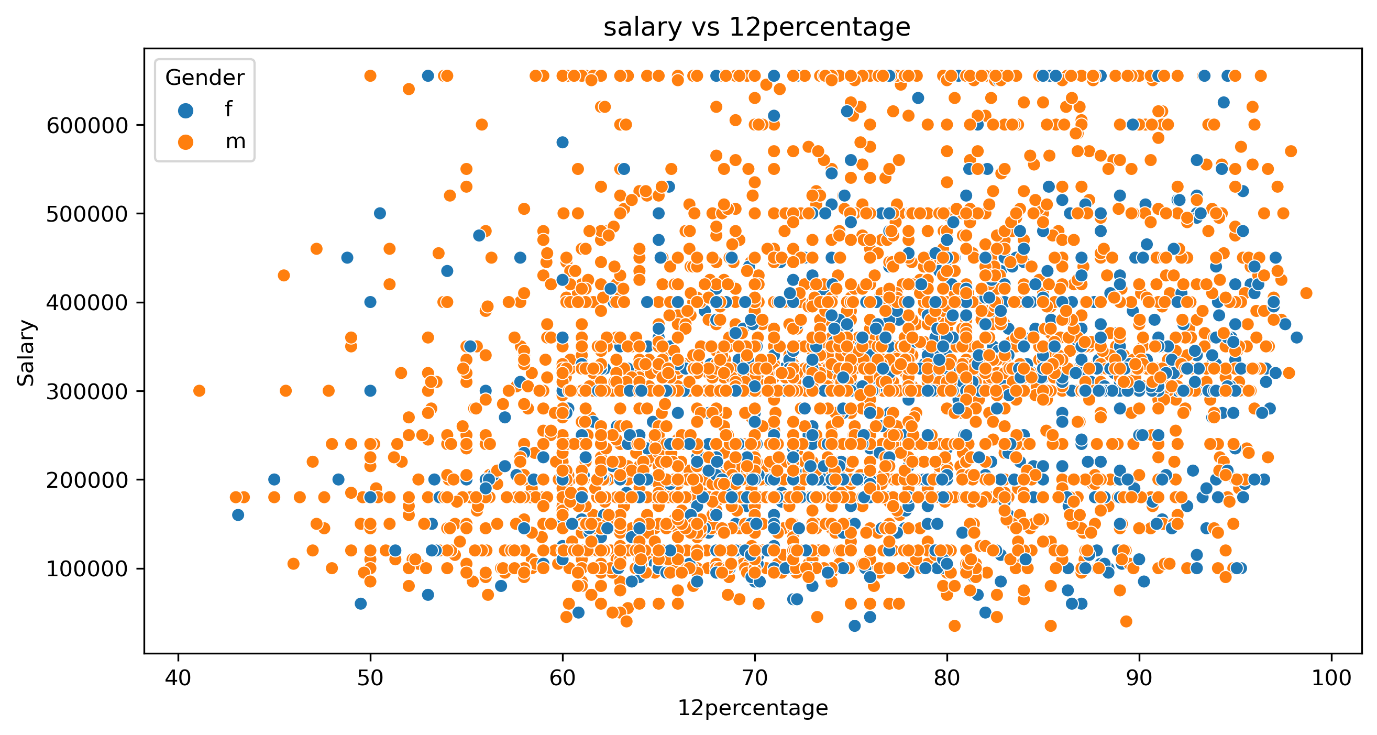
**Plot Type:** Scatter Plot

**Variables:**

* **X-axis:** 12percentage
* **Y-axis:** Salary
* **Grouping Variable:** Gender (Female, Male)

**Observations:**

* **Relationship between 12percentage and Salary:** There appears to be a weak positive correlation between 12percentage and salary. This suggests that, in general, individuals with higher 12percentage scores tend to have slightly higher salaries. However, the relationship is not strong, and there is a significant amount of scatter in the data, indicating that other factors also influence salary.
* **Gender-based Differences:** While there is a general trend of higher salaries associated with higher 12percentage scores for both genders, the scatter plot shows some overlap between male and female data points. This suggests that the relationship between 12percentage and salary may vary to some extent based on gender.

****

**Analysis of "Degree vs salary" Plot**

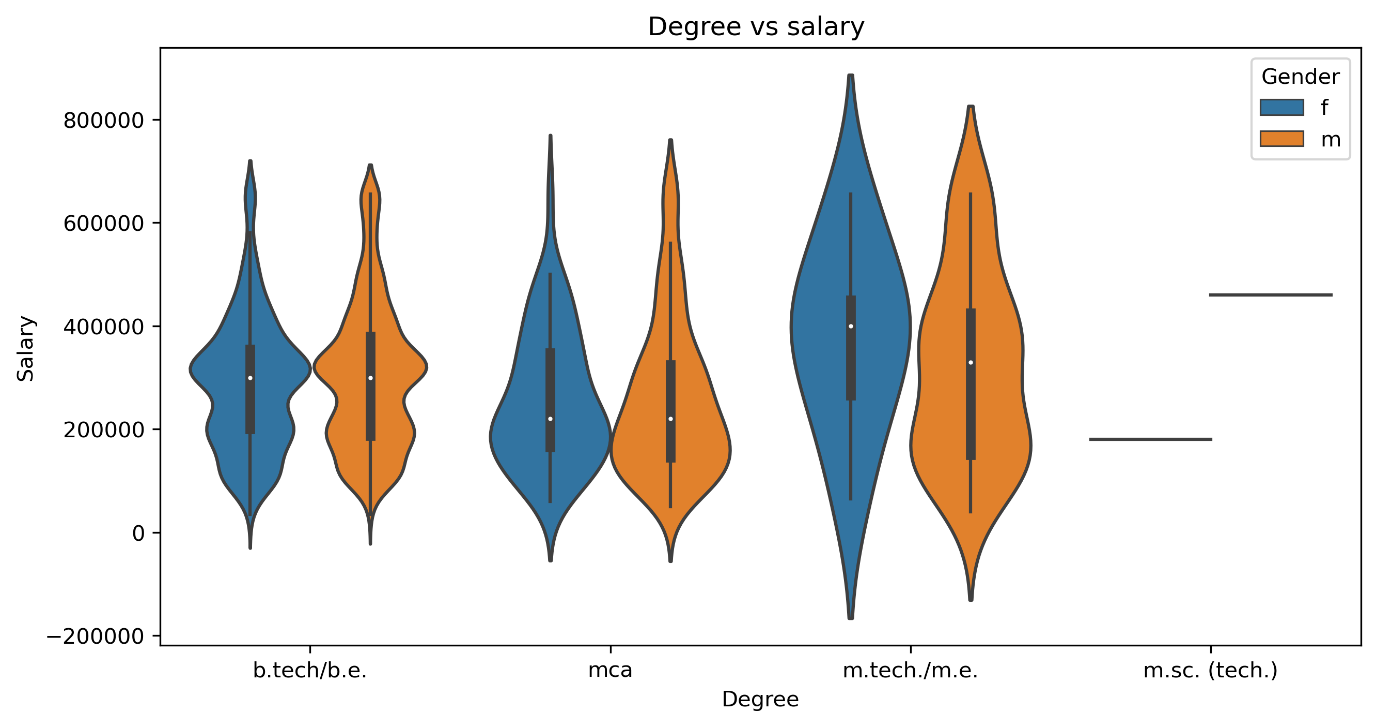
**Plot Type:** Violin Plot

**Variables:**

* **X-axis:** Degree (B.Tech/B.E., MCA, M.Tech/M.E., M.Sc. (Tech.))
* **Y-axis:** Salary
* **Grouping Variable:** Gender (Female, Male)

**Observations:**

* **Salary Distribution by Degree:** The violin plot provides a clear visualization of the salary distribution for each degree category and gender. The shape of the violin plots indicates the density of salaries within each group.
* **Degree-based Salary Differences:** The plot suggests that there are significant differences in salary distributions across the different degrees. For instance, M.Tech/M.E. graduates tend to have higher salaries compared to B.Tech/B.E. graduates, while MCA graduates' salaries are more spread out.
* **Gender-based Salary Differences:** Within each degree category, there are some variations in salary distributions between male and female graduates. In general, male graduates tend to have slightly higher salaries than female graduates, but the overlap in the violin plots indicates that there is also a significant degree of variation within each gender.

****

Analysis of "GraduationYear vs Salary" Plot

**Plot Type**: Scatter Plot

**Variables**:

\* X-axis: Graduation Year

\* Y-axis: Salary

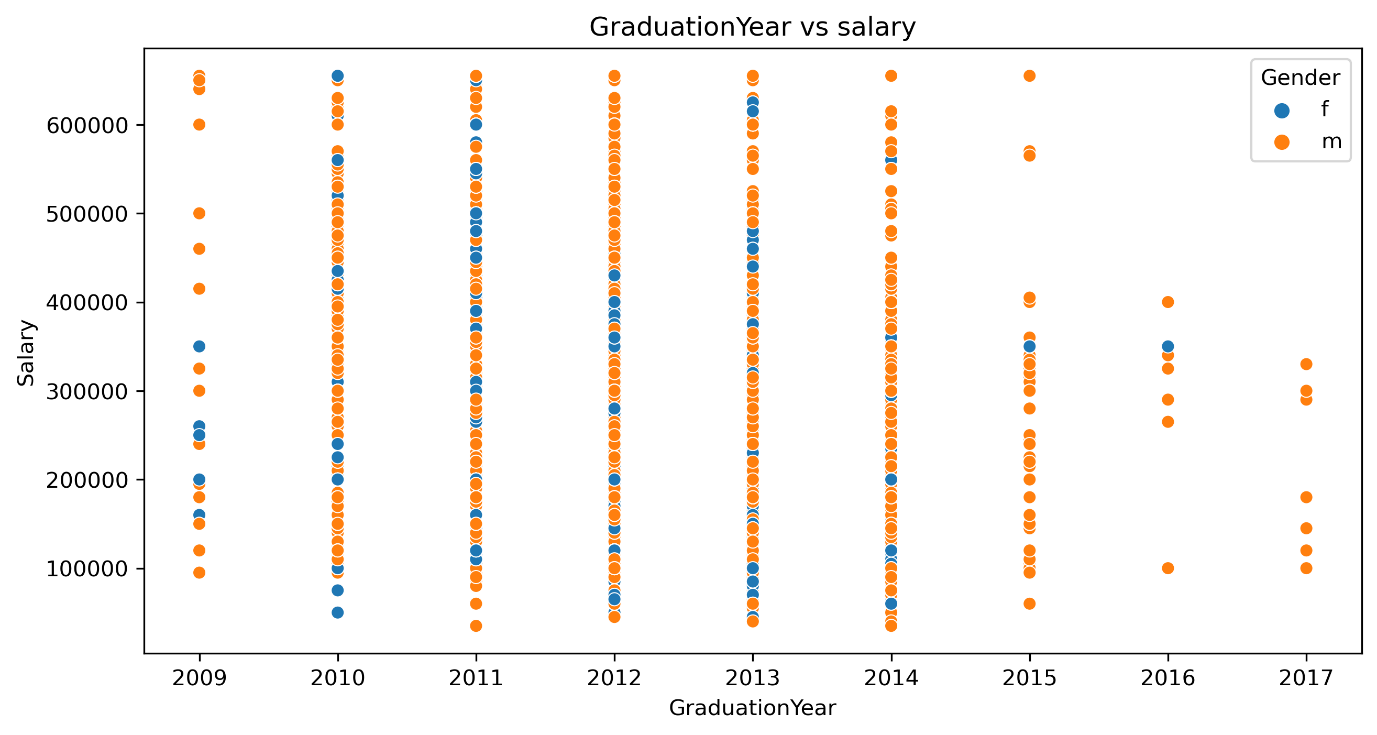
\* Grouping Variable: Gender (Female, Male)

**Observations:**

\* Salary Distribution by Graduation Year: The scatter plot shows the salary distribution over different graduation years (2009 to 2017). There are no clear patterns indicating that the graduation year directly correlates with a significant increase or decrease in salaries.

\* Salary Variation within Genders: Both male and female salaries are widely spread, with males having more data points, especially in the higher salary ranges. A slight trend of higher salaries in males can be observed across most graduation years.

\* Notable Outliers: Some individuals show significantly higher salaries regardless of gender, suggesting outliers or exceptional cases (possibly related to industry, role, or experience).



Analysis of "Salary vs CollegeGPA" Plot

**Plot Type:** Scatter Plot

**Variables:**

\* X-axis: College GPA

\* Y-axis: Salary

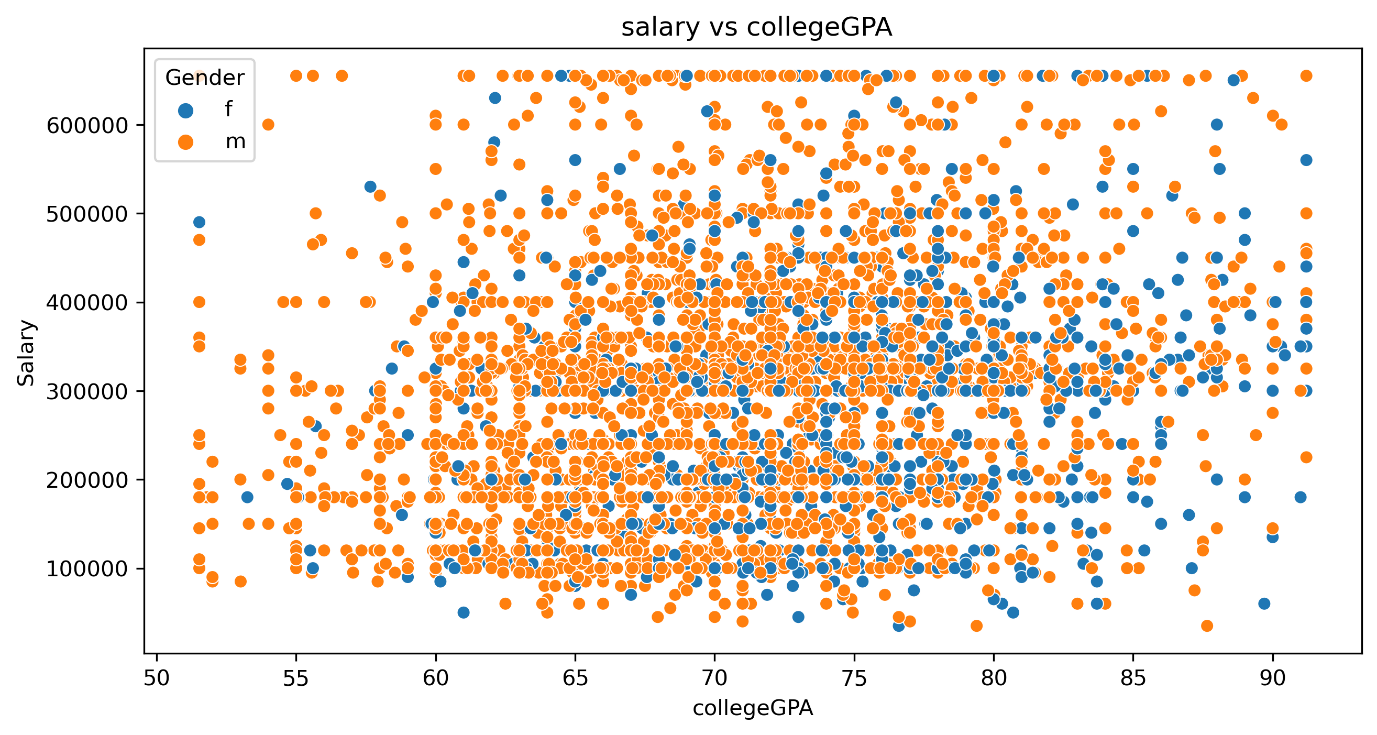
\* Grouping Variable: Gender (Female, Male)

**Observations:**

\* Salary vs GPA Trend: This plot indicates there is no strong correlation between college GPA and salary. Salaries are spread across all GPA ranges, and a high GPA does not consistently translate to a high salary.

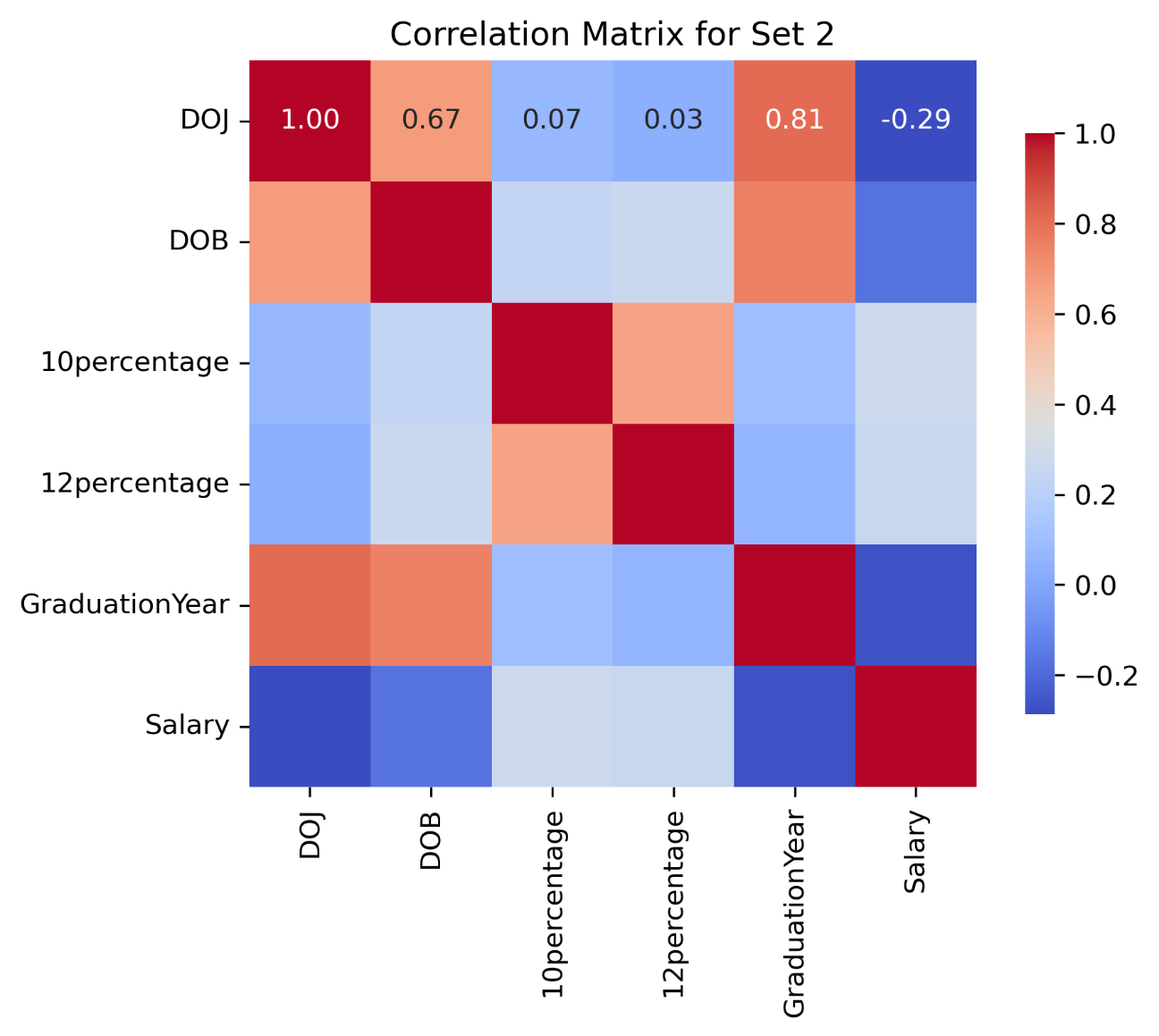
\* Gender-based Salary Differences: Both male and female graduates exhibit wide salary distributions, with a larger cluster of male graduates in the upper salary range. However, similar GPA ranges contain both high and low salary points for both genders, reinforcing the lack of clear GPA-salary correlation.

\* Outliers: A few outliers with very high salaries are present, mostly from the male population, though some females are also visible in these upper ranges.



Analysis of Correlation Matrix**Plot Type**: Heatmap**Variables:**

DOJ (Date of Joining), DOB (Date of Birth), 10percentage (10th-grade percentage), 12percentage (12th-grade percentage), GraduationYear, Salary**Observations:**\* DOJ and Salary Correlation: The correlation between DOJ (Date of Joining) and Salary is slightly negative (-0.29), indicating that more recent joiners may have lower salaries on average, which could be attributed to experience.\* DOB and Salary Correlation: A positive correlation between DOB and Salary (0.81) suggests that older employees (likely with more experience) tend to have higher salaries.\* School Grades and Salary: The 10th and 12th-grade percentages have low correlations with Salary, showing that school performance has little direct influence on salaries.\* Graduation Year and Salary: The negative correlation between GraduationYear and Salary (-0.29) indicates that recent graduates tend to earn lower salaries, which is expected due to their limited experience.Each plot gives insights into salary distribution across various factors, including graduation year, GPA, and gender. The correlation matrix helps further by providing insights into which factors have significant relationships with salary.



**General Observations:**

**1. \*Diagonal Elements (Histograms):**

- The diagonal elements show the distribution of individual variables.

- Variables like CollegeGPA and 12thPercentage seem to have more clustered distributions, while Salary has a wider range.

**2. Scatter Plots:**

- The scatter plots show pairwise relationships between variables. Below are some key observations for specific pairs:

**CollegeGPA vs Salary:**

- There seems to be a positive relationship between \*CollegeGPA\* and \*Salary\*. Higher GPAs are somewhat correlated with higher salaries, although the scatter plot shows a lot of spread.

**Graduation Year vs Salary:**

- Salary tends to be lower for those who graduated earlier (around 2010 or earlier), while those with more recent graduation years (2014-2016) tend to have higher salaries.

**12thPercentage vs CollegeGPA:**

- There appears to be a weak positive relationship, where students with higher percentages in 12th grade also tend to have higher College GPAs.

**Work Experience vs Salary:**

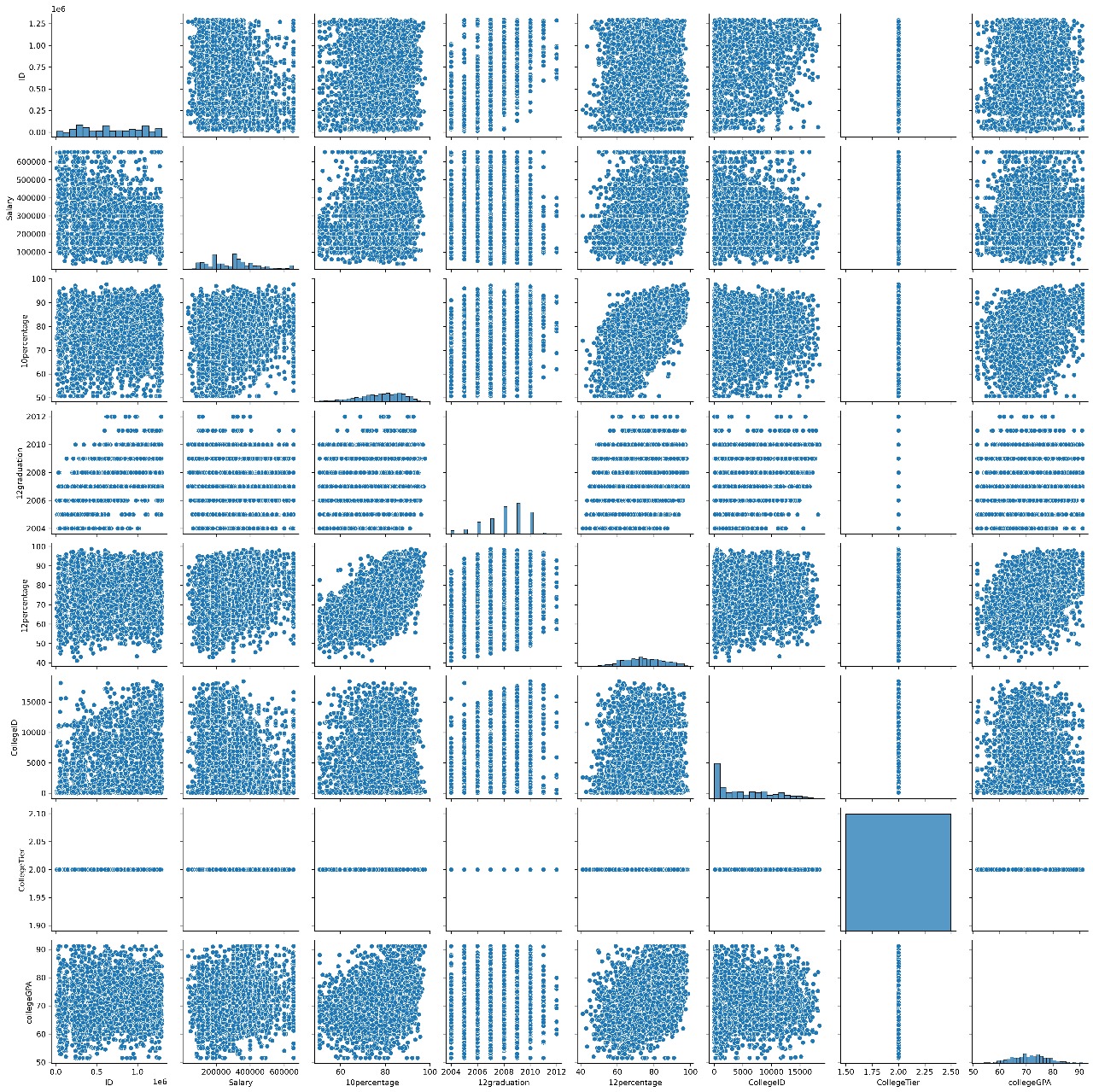
- There is a moderate positive relationship between \*Work Experience\* and \*Salary\*, indicating that as work experience increases, salaries also tend to increase.

**Clusters/Patterns:**

- The plots suggest some natural clusters or groupings, particularly in variables like \*Graduation Year\*, where there are distinct patterns based on the year of graduation.

**Outliers:**

- Some variables, especially \*Salary\* and \*12thPercentage\*, have potential outliers. These are points that are quite far from the general trend of the data.

General Observations:

**1. Diagonal Elements (Histograms):**

- The diagonal shows the distribution of variables like \*agreeableness, \*\*extraversion, \*\*neuroticism, \*\*openness to experience, and time-based data like \*\*DOJ (Date of Joining) year\* and \*DOB (Date of Birth)\*.

- Personality traits seem to have a normal distribution, centered around 0, with the majority of values falling between -2 and +2.

- \*DOJ year\* and \*DOB year\* have discrete distributions, indicating specific years for these events, rather than continuous variables.

**2. \*Scatter Plots:\***

- The scatter plots display pairwise relationships between personality traits and time-related variables:

**- \*Agreeableness vs Extraversion:\***

- There is no clear correlation between \*agreeableness\* and \*extraversion\*, as the data points seem to be spread randomly, forming a circular pattern.

**- \*Neuroticism vs Extraversion:\***

- Similarly, no strong relationship is observed between these two traits, suggesting that individuals’ scores on these traits do not directly influence one another.

**- \*Openness to Experience vs Agreeableness/Neuroticism:\***

- \*Openness to experience\* also shows no strong relationships with the other personality traits, as scatter plots seem evenly distributed without a clear trend.

**- \*DOJ Year vs Personality Traits:\***

- There is no apparent relationship between \*DOJ year\* and any of the personality traits. Individuals who joined in different years display a wide range of personality trait scores.

**- \*DOB Year vs Personality Traits:\***

- A similar observation is made for \*DOB year\* versus personality traits, indicating no significant relationship between birth year and personality trait distributions.

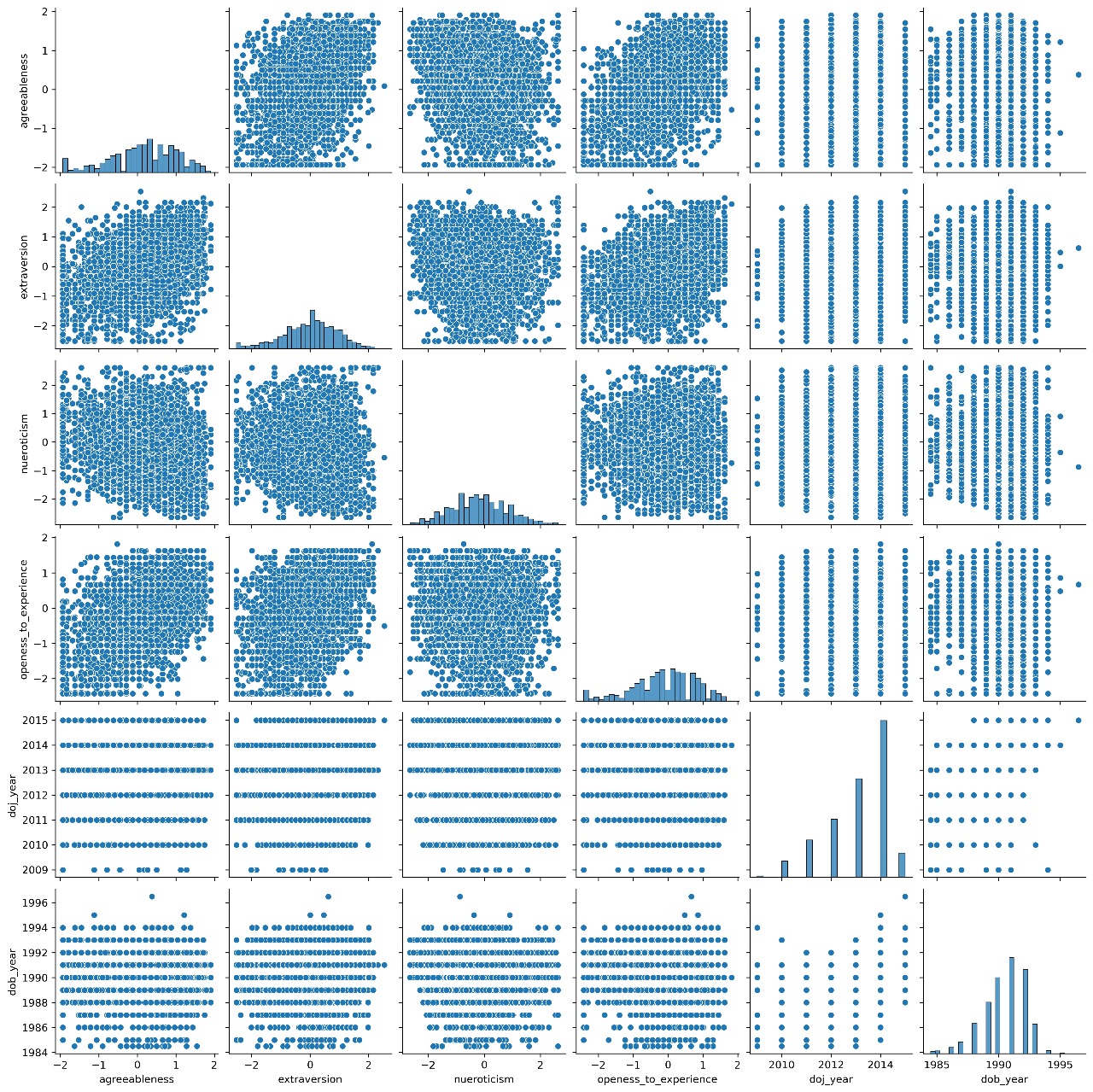
**3. \*Time-Based Patterns (DOJ Year and DOB Year):\***

- The \*DOJ year\* shows that a large number of individuals joined in the years between \*2010 and 2015\*. This may indicate a hiring trend or a dataset limited to a particular group of individuals who joined around this period.

- The \*DOB year\* distribution shows more concentration around \*1990-1995\*, which might suggest the majority of individuals in this dataset were born around these years.

**4. \*Clusters/Patterns:\***

- Personality traits do not show clear clustering based on time-based variables (like DOJ year or DOB year), indicating that individuals with different personality traits joined or were born in a wide range of years.



General Observations:

**1. \*Diagonal Elements (Histograms):\***

- The diagonal shows histograms for individual variables, giving insights into the distribution:

- \*Salary\* appears to be positively skewed, meaning most individuals have salaries concentrated at lower values with a few high earners.

- \*CollegeGPA\* and \*12thPercentage\* show tighter distributions, indicating that most students fall within a certain GPA range and similar high percentages.

- \*Work Experience\* is more spread out but seems to have some clustering around specific years.

- \*Graduation Year\* also has a clustered distribution, indicating most graduates fall within a particular range of years.

**2. \*Scatter Plots:\***

- The scatter plots show pairwise relationships between variables:

**- \*Salary vs Work Experience:\***

- There is a positive relationship here, as expected. More work experience generally correlates with higher salaries, though there is still significant scatter, meaning salary doesn't always increase at the same rate for everyone with more experience.

**- \*Salary vs CollegeGPA:\***

- There seems to be a slight positive relationship between \*CollegeGPA\* and \*Salary, though it’s weaker compared to \*\*Work Experience\*. This suggests that while GPA might impact salary, it's not the strongest determining factor.

**- \*Salary vs 12thPercentage:\***

- There is no clear trend between \*Salary\* and \*12thPercentage\*. This suggests that performance in earlier education (high school) does not have a strong direct relationship with salary in the current stage of these individuals' careers.

**- \*Work Experience vs CollegeGPA:\***

- There is no significant relationship between \*Work Experience\* and \*CollegeGPA\*, implying that people with different levels of GPA have similar amounts of work experience over time.

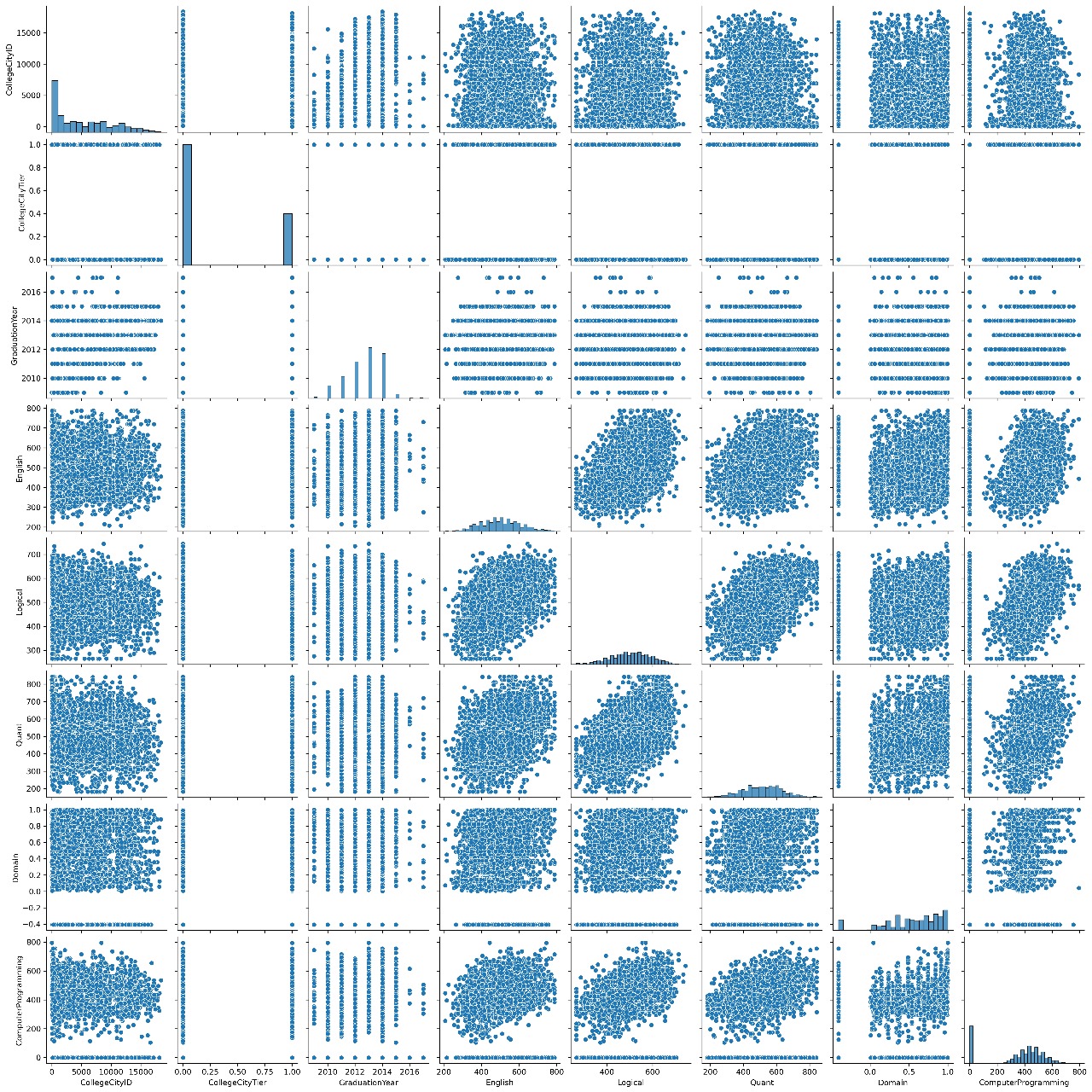
**- \*Graduation Year vs Salary:\***

- Individuals who graduated in more recent years (closer to 2010 and onwards) tend to have a wider range of salaries, indicating that more recent graduates may have more variance in salary outcomes, possibly due to different career paths or industries.

**3. \*Clusters/Patterns:\***

- The data shows clustering in \*Graduation Year\* and \*Work Experience\*, likely due to specific time periods where more individuals graduated or gained work experience, which might reflect the dataset’s sample distribution.

**4. \*Outliers:\***  There are potential outliers in the \*Salary\* data, where a few individuals earn significantly more than the rest. These could be people with exceptional experience or specialized roles.



\*Overall Salary Distribution by State:\*

- Salaries vary significantly by state. States like \*Delhi, Karnataka, and Telangana\* have higher median salaries, while states like \*Bihar, Assam, and Meghalaya\* have lower medians.

- The distribution of salaries across most states shows a wide range, as indicated by the length of the boxes (interquartile range) and the whiskers (spread of data).

**2. \*Gender-based Salary Differences:\***

- In most states, \*male salaries (orange)\* tend to be slightly higher than \*female salaries (blue). However, in states like \*\*West Bengal, Tamil Nadu, and Haryana\*, females seem to have equal or slightly higher median salaries.

- In states like \*Delhi, Karnataka, and Maharashtra\*, male salaries are visibly higher than female salaries, indicating a more significant gender gap in pay.

- In some states such as \*Bihar, Assam, Sikkim, and Meghalaya\*, we see fewer female data points, which could indicate a smaller sample size or fewer female professionals in those states.

**3. \*Outliers:\***

- Several states, including \*Delhi, Maharashtra, Karnataka, and \*\*Tamil Nadu, have visible outliers. These are individuals earning salaries much higher or lower than the general trend. For example, in \*\*Delhi\*, there are high-earning outliers for both males and females.

- \*Goa\* and \*Sikkim\* show larger spreads, where a few individuals are earning considerably higher or lower salaries compared to the majority.

**4. \*States with Balanced Distribution:\***

- In states like \*Kerala, West Bengal, and Tamil Nadu\*, the salary distribution between genders seems more balanced, with smaller differences in median salaries and overlapping salary ranges.

**5. \*States with Wide Gender Gaps:\***

- States like \*Madhya Pradesh, Gujarat, and Rajasthan\* exhibit a noticeable salary difference between males and females, where males typically have higher earnings.

Insights:

- \*Higher-paying states: \*\*Delhi, Karnataka, Maharashtra, and \*\*Telangana\* generally offer higher median salaries, which could be attributed to the presence of larger industries or job markets.

- \*Lower-paying states: \*\*Bihar, Assam, and Meghalaya\* show lower median salaries, potentially due to a less developed job market or different economic factors.

- \*Gender gap\*: In most states, males earn more than females, but in a few states, the gender gap appears narrower or even reversed.



**Engineering Specializations VS Salary**

The specialization that stands out the most is "mechatronics" with the highest salary, significantly above 600,000.

Specializations like "chemical engineering" and "computer science" also have relatively higher salaries compared to others.

Specializations with Lower Salaries:

"Textile technology," "biomedical engineering," and "electronics and telecommunication" are among the specializations with the lowest salaries, hovering around or slightly below 200,000.

Salary Range:

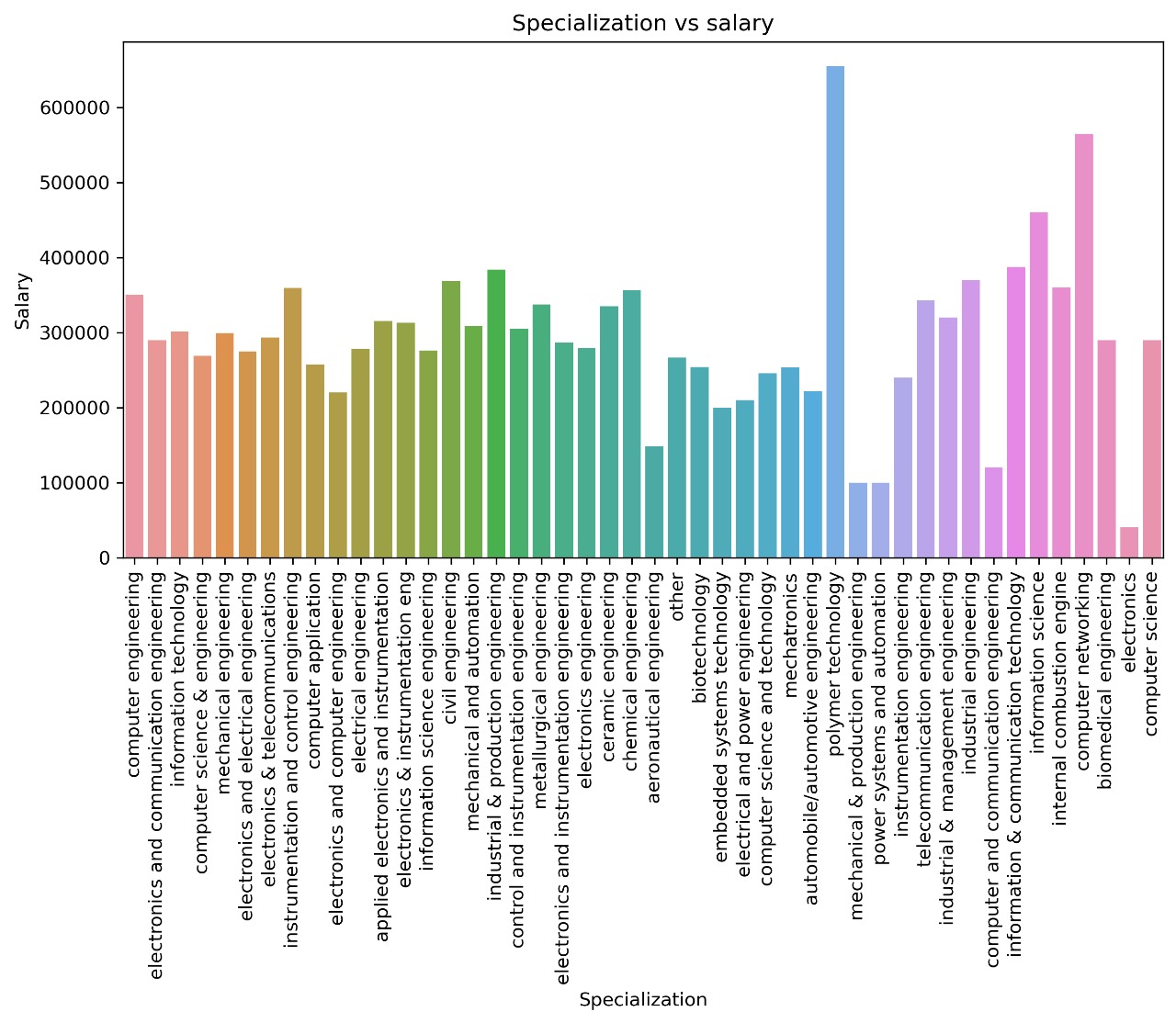
The salaries vary widely, with the highest being above 600,000 and the lowest just under 200,000.

Most specializations have salaries in the range of 200,000 to 400,000.

Notable Patterns:

Some fields, particularly those related to newer or interdisciplinary areas such as "mechatronics" and "biotechnology," show higher salaries compared to traditional engineering disciplines like "mechanical engineering" and "civil engineering."

The chart provides a useful overview of salary variations across different specializations, which could help guide career decisions based on financial outcomes in different fields.



**Conclusion:**

**The analysis of the AMCAT dataset reveals several insights into salary trends, correlations with educational background, gender-based differences, and other factors influencing compensation. Key conclusions include:**

**1. \*Salary Influences\*:**

**- There's no strong correlation between GPA, high school scores, or graduation year with salary, though experience plays a more significant role.**

**- Some specializations, such as Mechatronics and Computer Science, offer higher salaries, while others like Textile Technology and Biomedical Engineering tend to have lower salaries.**

**2. \*Gender Pay Gap\*:**

**- Male graduates generally earn more than female graduates, although the gap varies by year, degree, and state.**

**3. \*Experience and Salary\*:**

**- A moderate positive correlation exists between work experience and salary, with more experienced employees tending to earn higher salaries.**

**4. \*Geographical Variations\*:**

**- Salaries are higher in states like Delhi and Karnataka, while states like Bihar and Assam show lower salary medians.**

**5. \*Outliers\*:**

**- Several outliers, particularly in salary, highlight exceptional cases potentially linked to roles or individual experience.**