Assignment 03

Soham Deshkhaire

October 29, 2025

# 1. Companies Table

|  | company\_name | company\_raw | company\_is\_staffing | company\_id |
| --- | --- | --- | --- | --- |
| 0 | Crowe | Crowe | False | 0 |
| 1 | The Devereux Foundation | The Devereux Foundation | False | 1 |
| 2 | Elder Research | Elder Research | False | 2 |
| 3 | NTT DATA | NTT DATA Inc | False | 3 |
| 4 | Frederick National Laboratory For Cancer Research | Frederick National Laboratory for Cancer Research | False | 4 |

# 2. Data Preparation (Clean Up Data)

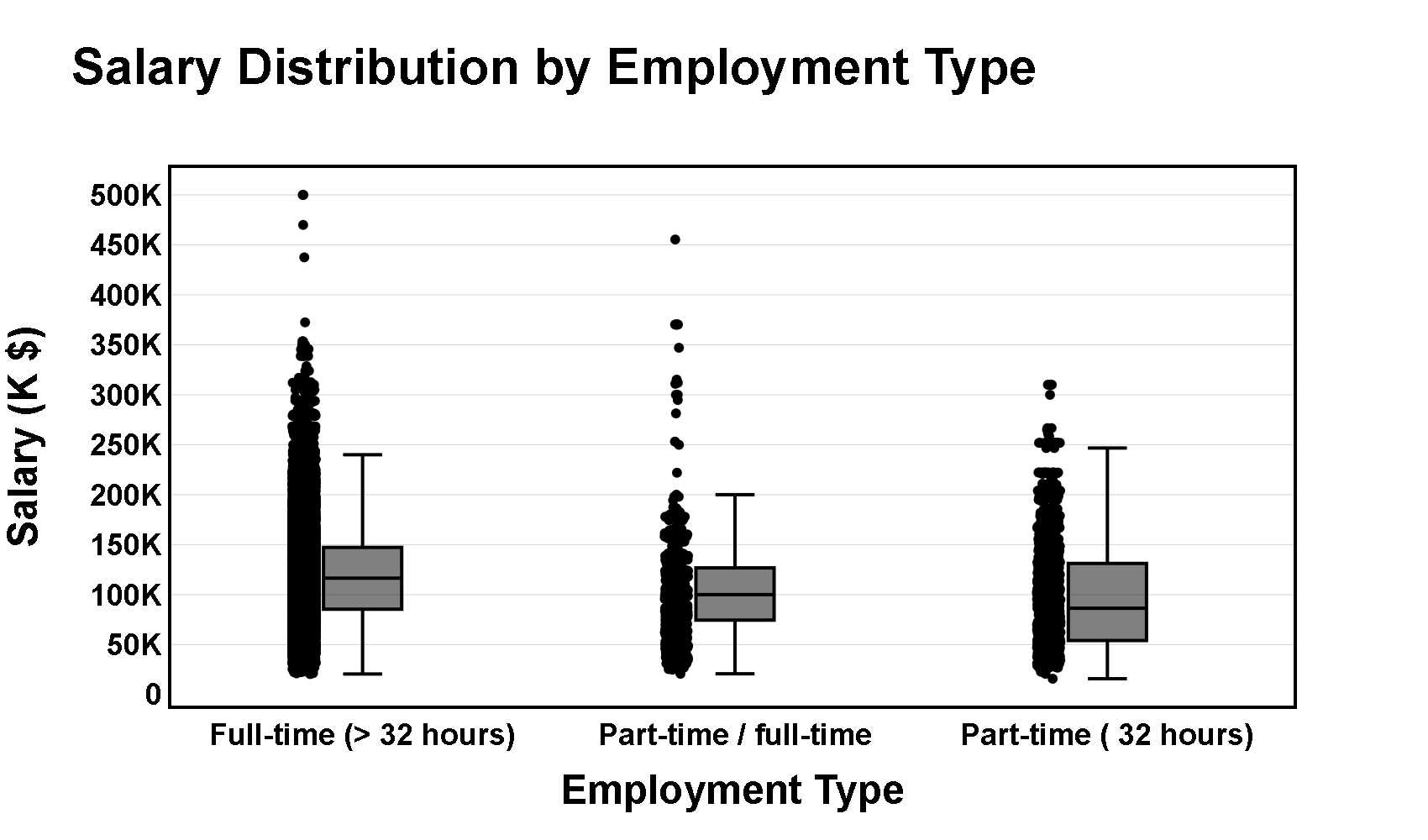
Medians : 87295.0 130042.0 115024.0  
Data cleaning complete. Rows retained: 72498

|  | EMPLOYMENT\_TYPE\_NAME | SALARY |
| --- | --- | --- |
| 0 | Part-time / full-time | 92500.0 |
| 1 | Full-time (> 32 hours) | 110155.0 |
| 2 | Full-time (> 32 hours) | 92962.0 |
| 3 | Full-time (> 32 hours) | 107645.0 |
| 4 | Full-time (> 32 hours) | 192800.0 |

|  | EMPLOYMENT\_TYPE\_NAME | SALARY |
| --- | --- | --- |
| 0 | Part-time / full-time | 92500.0 |
| 1 | Full-time (> 32 hours) | 110155.0 |
| 2 | Full-time (> 32 hours) | 92962.0 |
| 3 | Full-time (> 32 hours) | 107645.0 |
| 4 | Full-time (> 32 hours) | 192800.0 |

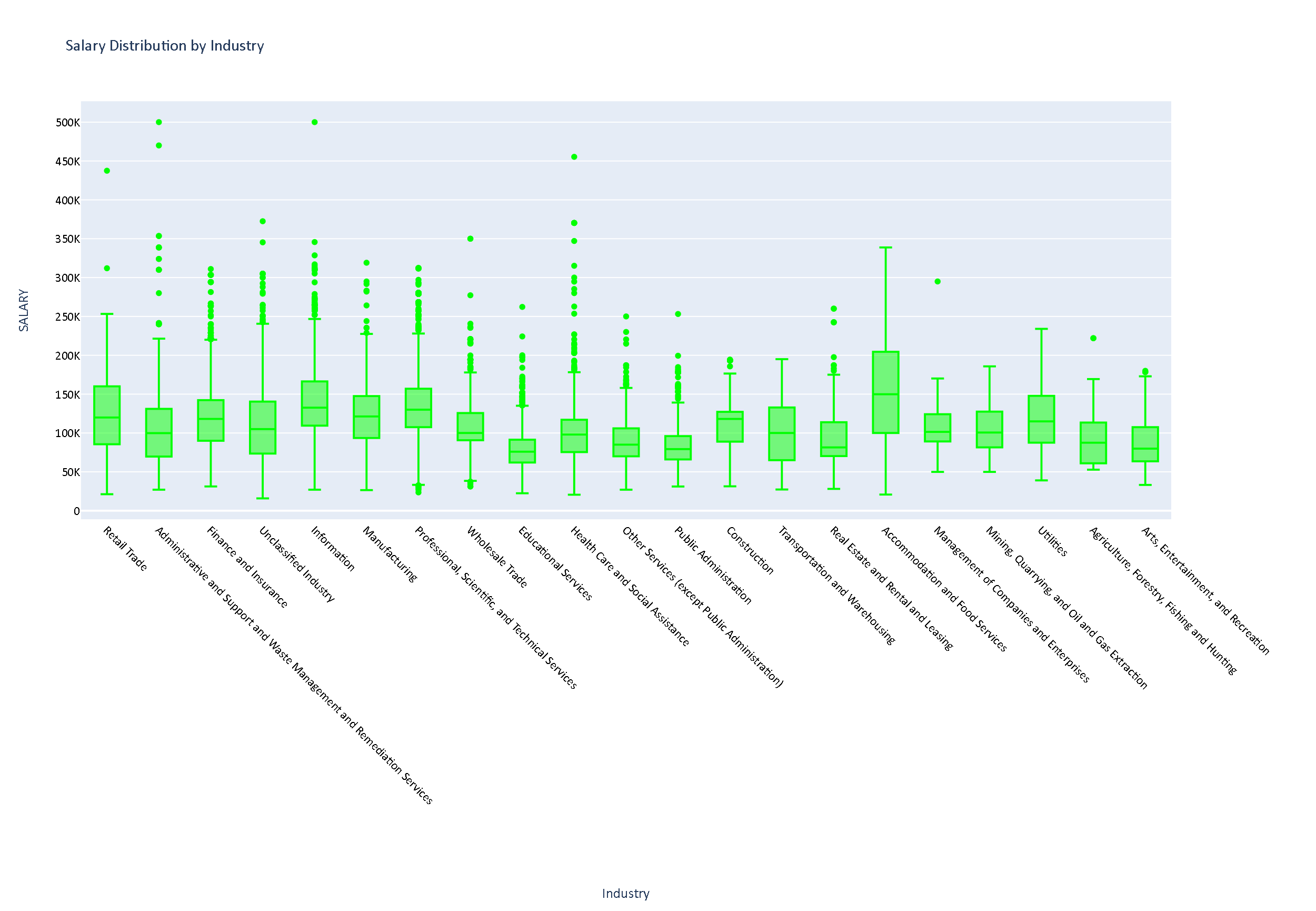
EMPLOYMENT\_TYPE\_NAME  
Full-time (> 32 hours) 116500.0  
Part-time ( 32 hours) 86390.0  
Part-time / full-time 100000.0  
Name: SALARY, dtype: float32

EMPLOYMENT\_TYPE\_NAME  
Full-time (> 32 hours) 116500.0  
Part-time ( 32 hours) 86390.0  
Part-time / full-time 100000.0  
Name: SALARY, dtype: float32



# 3. Salary Distribution by Industry and Employment Type

Compare salary variations across industries.  
Filter the dataset  
 Remove records where salary is missing or zero.  
Aggregate Data  
 Group by NAICS industry codes.  
 Group by employment type and compute salary distribution.  
Visualize results  
 Create a box plot where:  
 X-axis = NAICS2\_NAME  
 Y-axis = SALARY\_FROM  
 Group by EMPLOYMENT\_TYPE\_NAME.  
 Customize colors, fonts, and styles.  
Explanation: Write two sentences about what the graph reveals.

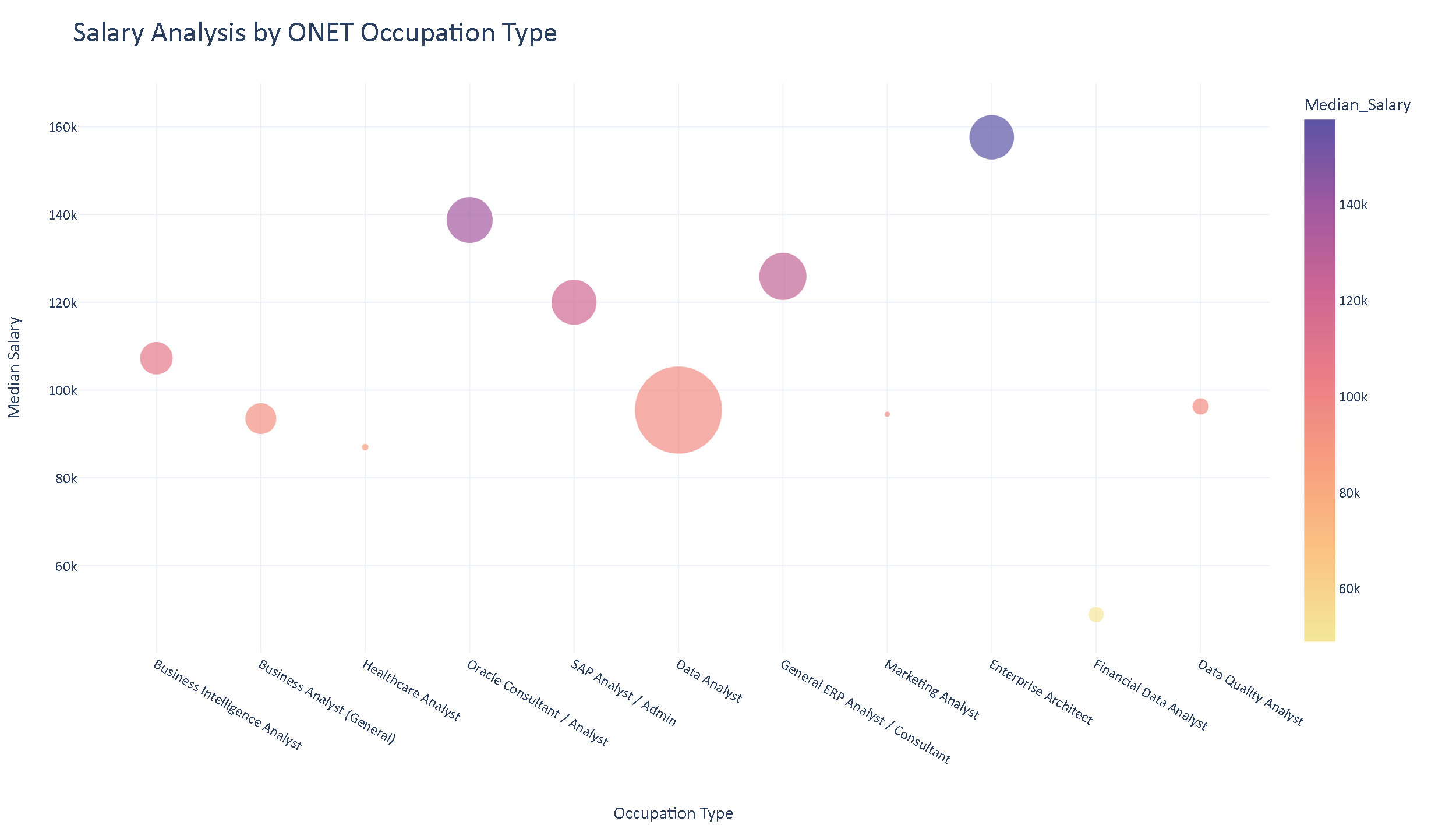


# This graph highlights both the highest and lowest ends of job salaries, clearly showing outliers across various industries. It also points out the wide gap in accommodation and food services, where median salaries range from around $100K to nearly $200K, compared to sectors like education services with medians closer to $60K–$80K.

# 4. 3 Salary Analysis by ONET Occupation Type (Bubble Chart)

–Appendix 1: I asked Copilot for assistance because my aggregation wasn’t working properly. It turned out the issue was caused by a combination of the aggregation and sorting steps we had implemented during the Saturday help session, as shown in the attached AI prompts.

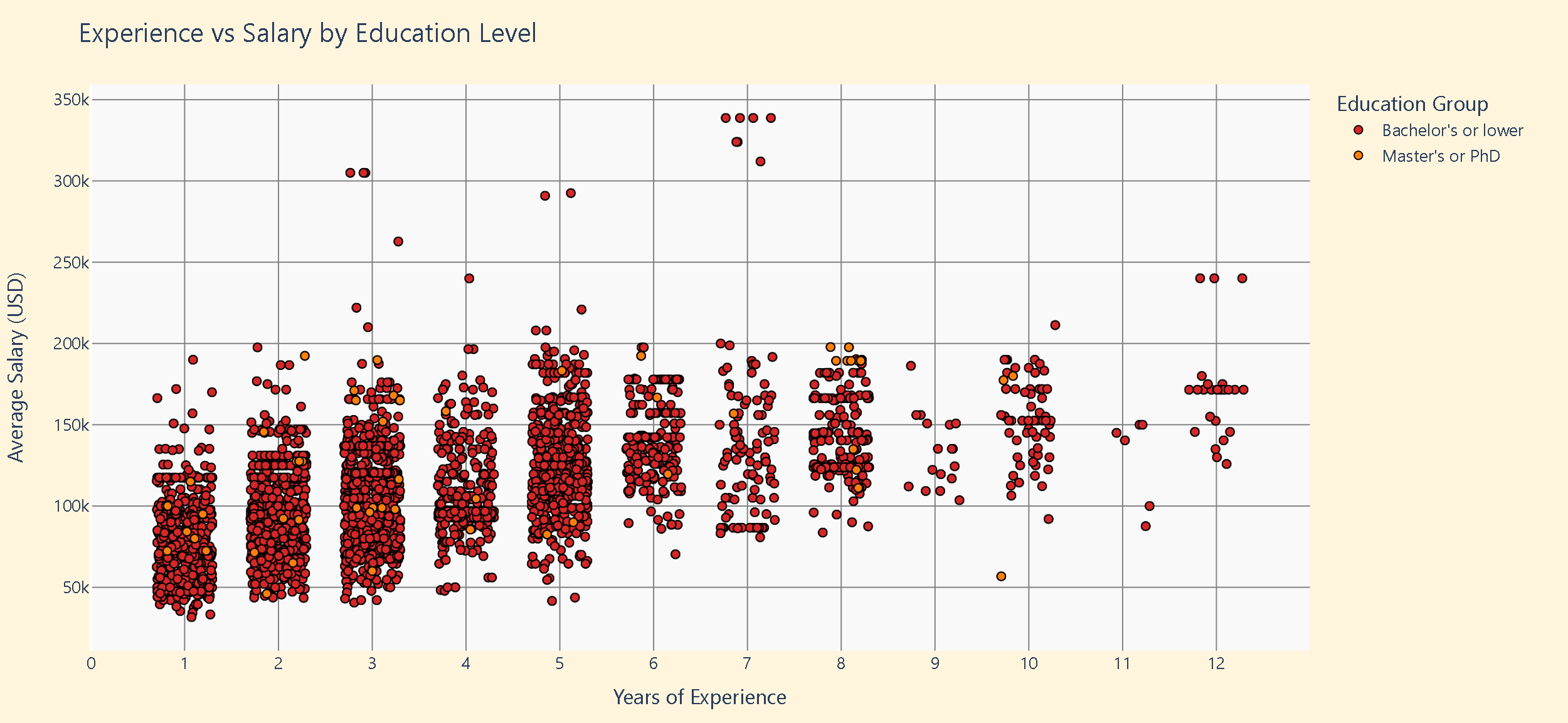
Analyze how salaries differ across ONET occupation types.  
Aggregate Data  
 Compute median salary for each occupation in the ONET taxonomy.  
Visualize results  
 Create a bubble chart where:  
 X-axis = ONET\_NAME  
 Y-axis = Median Salary  
 Size = Number of job postings  
 Apply custom colors and font styles.  
Explanation: Write two sentences about what the graph reveals.



# This graph shows that while data analysts represent the largest field with the most data entries, enterprise architects earn the highest average salaries in this dataset. It also indicates that specialized roles requiring niche skills tend to pay more than broader positions like business analysts, though business intelligence analysts still stand out with a comparatively higher average income.

# 5. 4 Salary by Education Level (Two Groups)

Create two groups:  
 Associate’s or lower (GED, Associate, No Education Listed)  
 Bachelor’s (Bachelor’s degree)  
 Master’s (Master’s degree)  
 PhD (PhD, Doctorate, professional degree)  
Plot scatter plots for each group using, MAX\_YEARS\_EXPERIENCE (with jitter), Average\_Salary, LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME  
After each graph, add a short explanation of key insights.

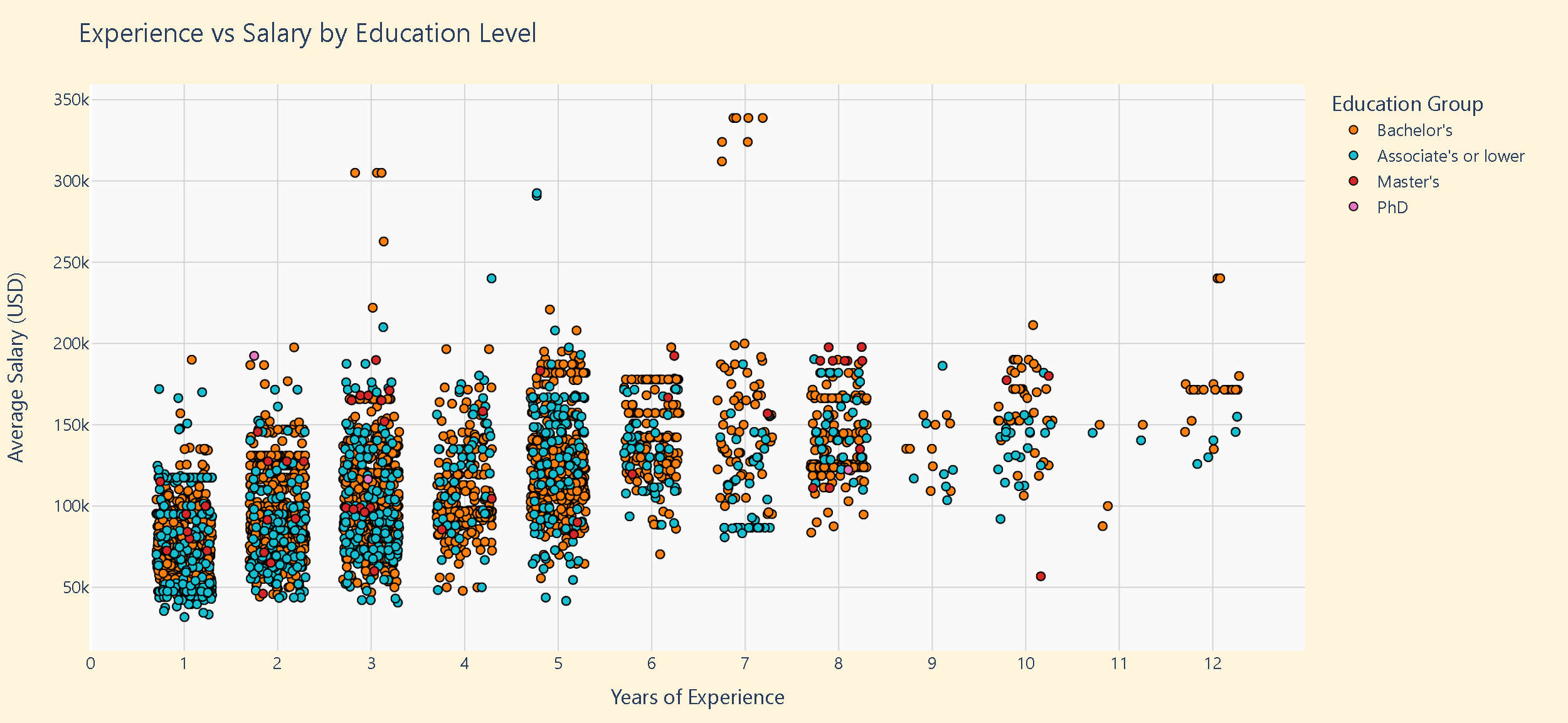


# This graph suggests that earning a master’s degree can lead to higher starting salaries, often compensating for fewer years of experience, and tends to result in higher average pay over time. However, there are exceptions—such as master’s holders with around 10 years of experience earning roughly $60K—indicating that factors like company requirements or role type may also influence salary outcomes.

# 6. 4 Salary by Education Level (Four Groups)

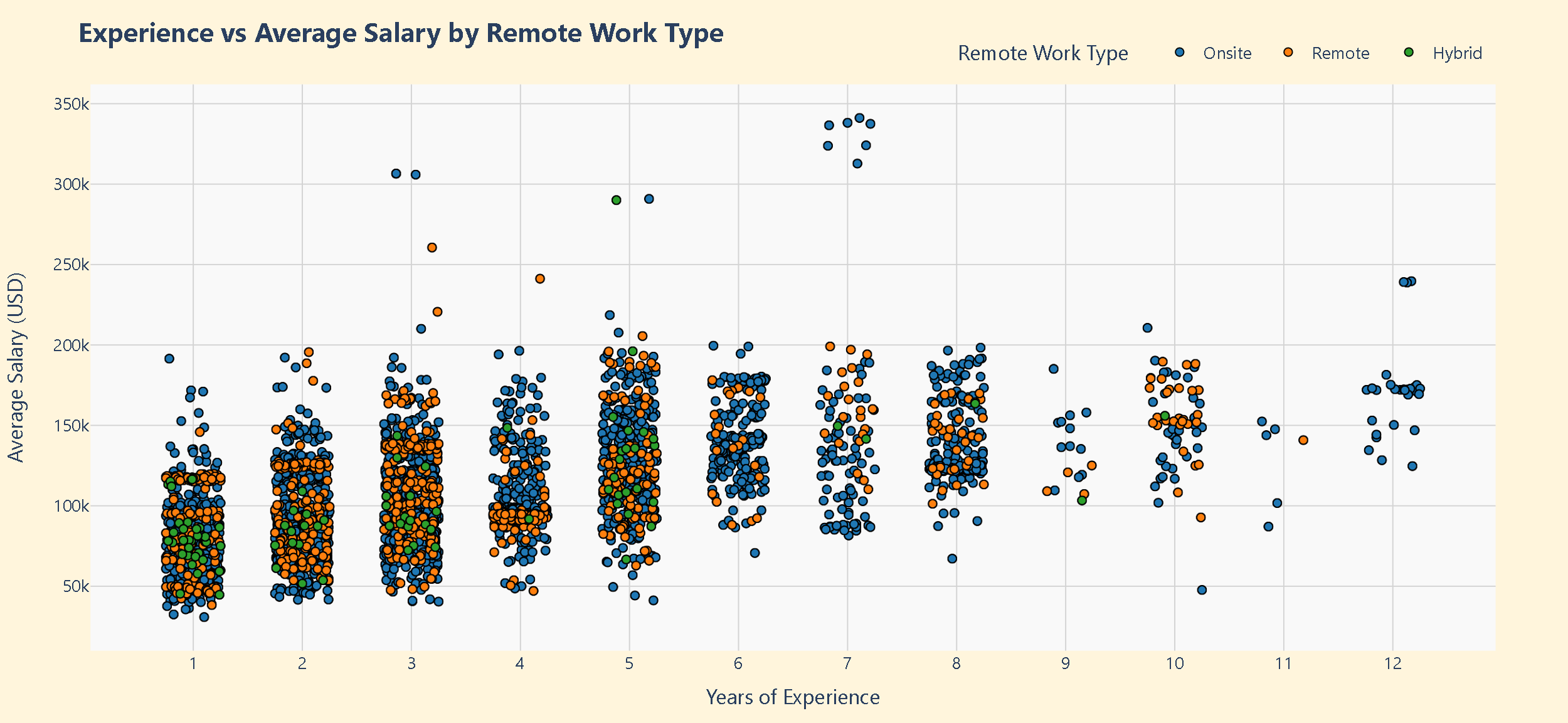
Create two groups:  
 Associate’s or lower (GED, Associate, No Education Listed)  
 Bachelor’s (Bachelor’s degree)  
 Master’s (Master’s degree)  
 PhD (PhD, Doctorate, professional degree)  
Plot scatter plots for each group using, MAX\_YEARS\_EXPERIENCE (with jitter), Average\_Salary, LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME  
After each graph, add a short explanation of key insights.

See appendix 2 – asked ai to help me fix the data being in a straight line and it suggested the jitter.



This graph shows that while PhD holders may start their careers with higher salaries, over time, years of experience become the dominant factor influencing pay. The data suggests that individuals with associate or bachelor’s degrees often see greater salary growth later in their careers compared to those with higher-level degrees.

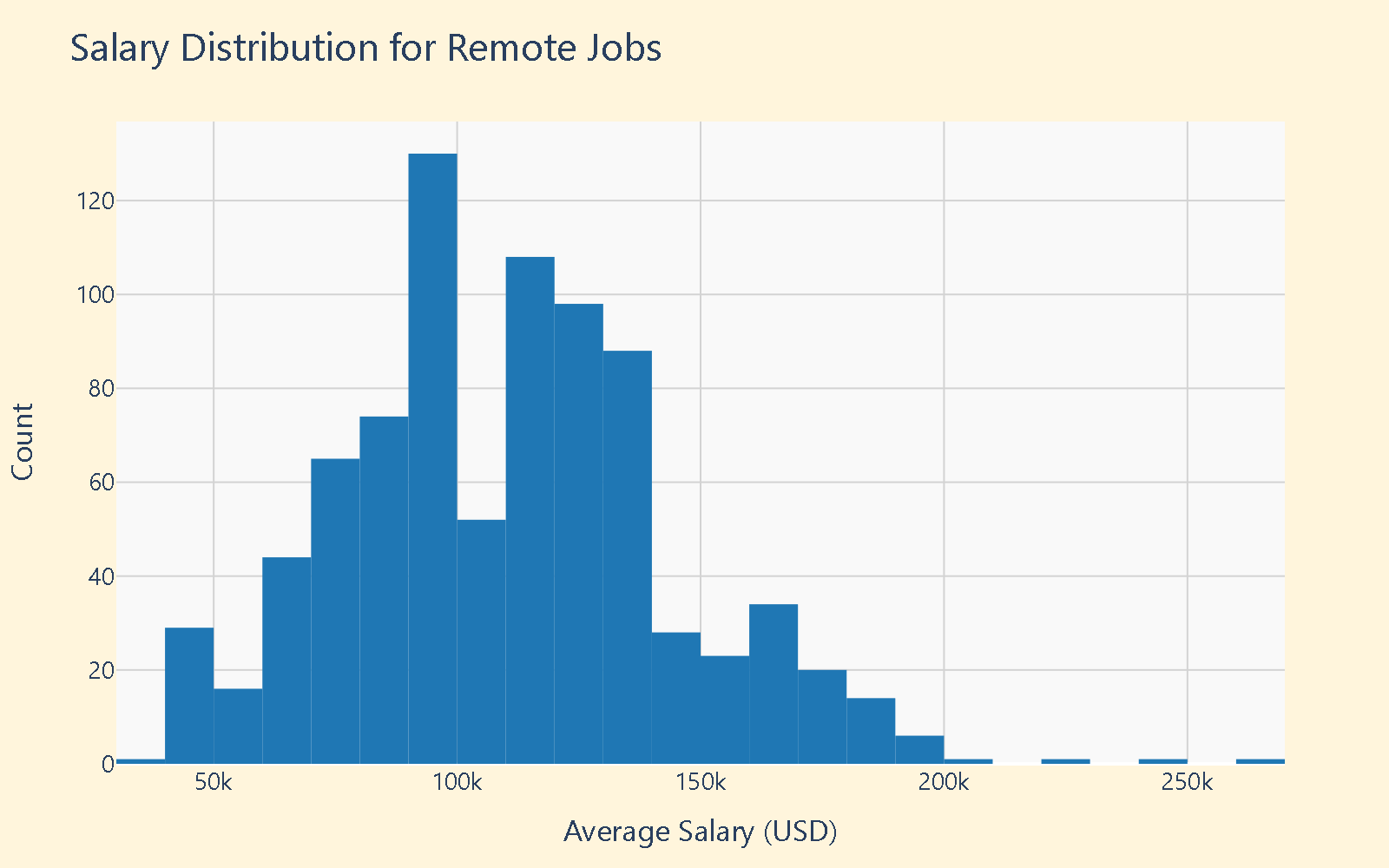
|  | MAX\_YEARS\_EXPERIENCE | Average\_Salary | LOT\_V6\_SPECIALIZED\_OCCUPATION\_NAME | REMOTE\_GROUP |
| --- | --- | --- | --- | --- |
| 0 | 2.0 | 92962.0 | Data Analyst | Onsite |
| 1 | 2.0 | 75026.0 | Oracle Consultant / Analyst | Onsite |
| 2 | 1.0 | 60923.0 | Data Analyst | Remote |
| 3 | 2.0 | 131100.0 | Enterprise Architect | Onsite |
| 4 | 3.0 | 136950.0 | Enterprise Architect | Remote |

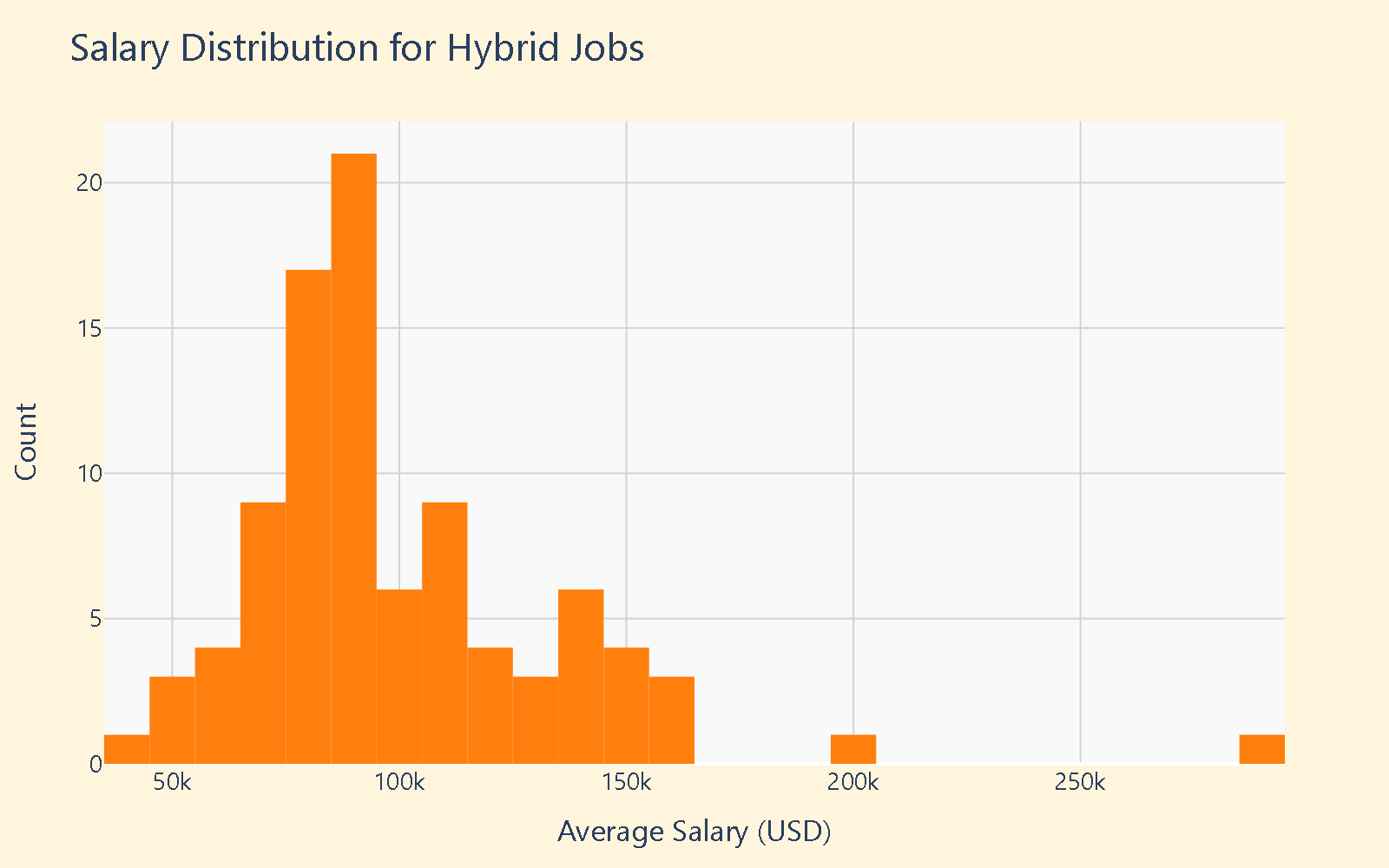


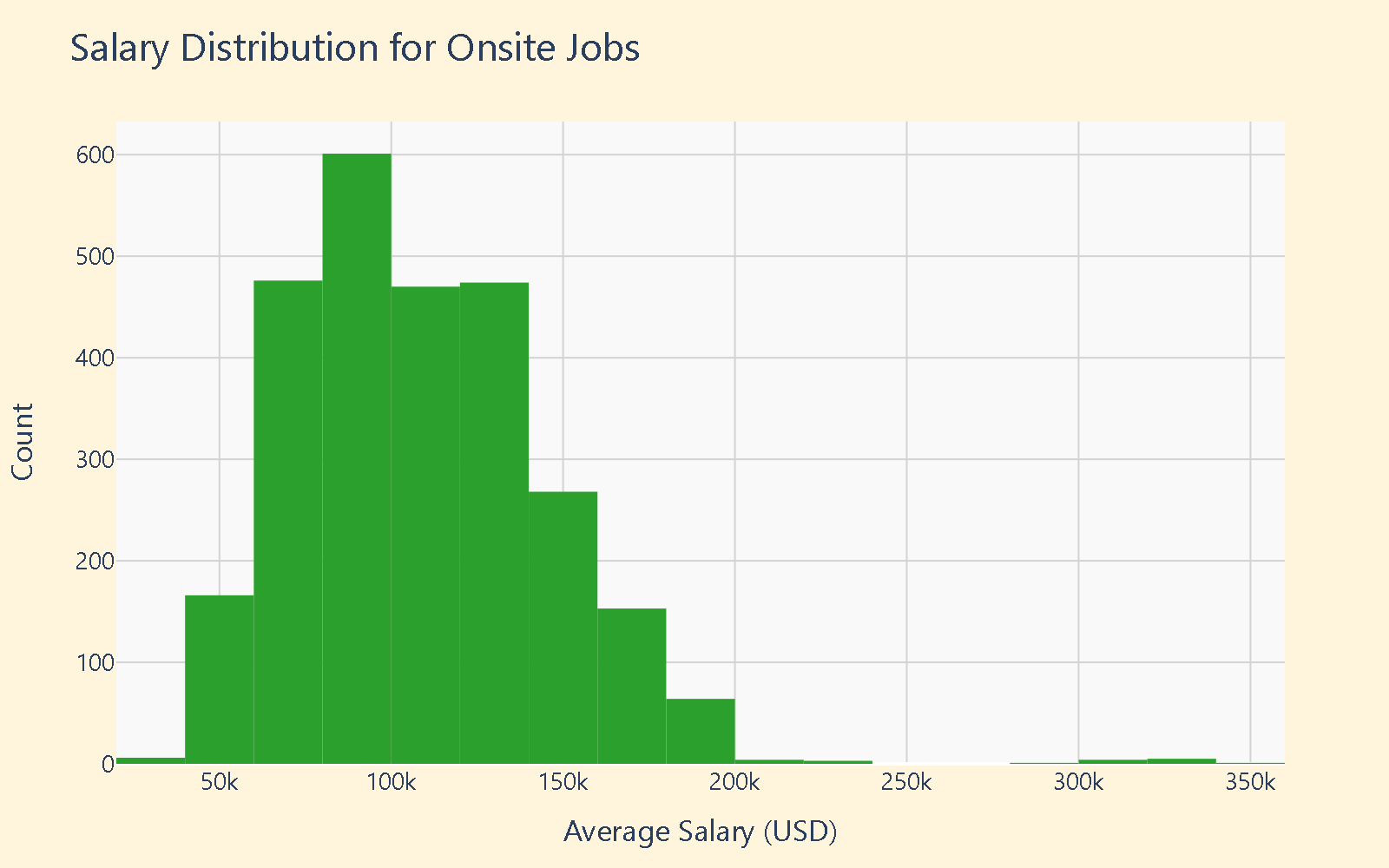
# This graph shows that there isn’t a strong link between years of experience and salary, nor a clear difference in pay between remote and onsite roles. Salaries appear to peak around the seven-year mark, but overall, the data suggests fairly balanced compensation across remote, onsite, and hybrid positions.

# 7. Histrograms

# 8. Reference Appendix 3 for AI Prompt, had data but needed help formulating code to get all histographs in one.







**Remote Jobs** – The data shows fewer low-paying remote positions, suggesting that these roles often require higher skills or more experience, which may explain the flexibility in work location and the concentration of salaries around the $100K range. However, there are also relatively few very high-paying remote roles, indicating a narrower distribution at the top end.

**Hybrid Jobs** – A similar pattern appears here but shifted slightly lower, with most salaries clustering around $80K. The sample size is smaller—roughly 25 jobs in each top salary bracket—so while trends are visible, they’re based on less data than for remote or onsite roles.

**Onsite Jobs** – Onsite positions make up the majority of the dataset, showing a more evenly distributed salary range that gradually declines toward the $200K mark. Unlike the other categories, there isn’t a sharp spike in job count at any specific salary level, reflecting broader variability due to the larger data volume.