Project Report

# **GitHub URL**

<https://github.com/Kattyn9/UCDPA_Katrina_Nee>

# **Abstract**

In this project I look at Arlington County employee demographics and how they relate to pay. From this analysis I hope to identify variations in salary across race and gender and potential causes of these variations. Using coding methods and functions I hoped to gain insights from the data.

# **Introduction**

I tried to find a data set to use for this project that would mirror data I work with on a regular basis. I work in the marketing department of a pension and finance brokers. As a part of my role I am often looking into the demographics of public sector employees as they are a large part of our client base. I thought it would be interesting to look to the salaries and demographics of public sector employees in America. A large part of my current role is being able to segment large amounts of data into smaller demographic and employer sections, in order to produce targeted marketing campaigns.

# **Dataset**

The first dataset I chose was of Arlington county, VA employee demographics (Employee Demographics, 2020). I thought this would be an interesting base to start from as the dataset included demographic data.

As I worked through the code I found it would be interesting to merge another dataset into my project. I was able to find a PowerPoint by Arlington county government that have the pay scales for county employees (Fiscal Year 2023 Pay Schedule for County Employees, slides 3 - 33, 2023). This dataset provided the pay scale information that matched with the job titles of the county employees from my first dataset.

# **Implementation Process**

I first imported pandas onto my Jupyter notebook, and proceeded to bring in my dataset of the employee demographic data from Arlington, Virginia (Data.gov, 2020). I then wanted to know how this data was structured so I ran “Employee.info()”, I then proceeded to sort the data by job title and in ascending order.

I then thought it would be helpful for future coding if I was to change the job title into the dataset’s index, I did this by calling “Employee.set\_index('JobTitle').head(2)”. I was curious to see what the average start year was for each gender, so I ran a groupby function, I found that the average start year was 2007 for Female and 2006 for Male.

I then proceeded to use the groupby function to group by job title, race and sex. In trying to remove the duplicates from the data, I encountered an issue, there was no way to distinguish each of the employees from one another, such as a Social Security Number, this meant that by dropping the duplicates I was left with only 5 lines of data. I opted to not drop the duplicates as I felt this would be compromising the dataset. I also encountered issues when trying to replace missing values as each time I ran the function with different criteria it returned that there were no missing values in the dataset. In a markdown line I decided to exhibit had there been missing values the code I would have used to remove and/or replace them.

It was at this point that I looked for another dataset that could help expand on the data I had and help me gain more of an insight into the county employees of Arlington. I was able to find pay scale data for Arlington County employees from July 2022 to June 2023 (Fiscal Year 2023 Pay Schedule For County Employees. 2022). The data was in the form of a PowerPoint that was available as a pdf. I was able to transfer the data I needed into an excel and save as a CSV file. I removed the data from other slides in the PowerPoint that would not have been relevant to the dataset I was examining. Once I had the data converted into a CSV file, I imported the data into my coding on Jupyter.

I applied the pd.read\_csv(‘Arlington County Employee Salary.csv’), this allowed me to see how the data transferred over and what each column contained. I then set the name of this data to Ar\_Sal as it would be easier to call into later code. I ran the info function to gain an insight into how the data was arranged and structured.

After this I sorted by job title and in ascending order so that it would be in the same order as the Employee dataset from earlier. I changed the index to be the job title, this way both datasets are arranged in the same manner. Before merging the tables I wanted to compare the two datasets and see how they aligned and if I could prevent any issues occurring when I went to merge them. I decided to merge them using the job title as a common factor between the two datasets, this would allow the data from the Ar\_Sal data set to be merged into the Employee one with out corrupting the data. I then ran the code to merge the datasets.

Once merged I wanted to run some conditional statements on the data. Before doing this I sorted the new merged dataset, Employee\_sal, by the employee’s start year. I used a loop with an if, elif and else function then that once input with an employee’s start year it would return the likely annual salary bracket. This would allow for a quick look up of the salary range of the employees based on their start year working for Arlington county.

I proceeded to apply a For Loop via iterrows on the Employee\_sal dataset. After this I wanted to look more closely at the demographics data in relation to the hourly maximum salary and if there was patterns to how different demographics of employees are paid. To do this I ran a groupby function looking at the race, sex and then the average hourly maximum salary. I then looked at the proportions of each race and gender of the employees, using a count values function. I created a reusable function that when the max and min of an employees salary was input would return the range of hourly pay that employee is likely to receive.

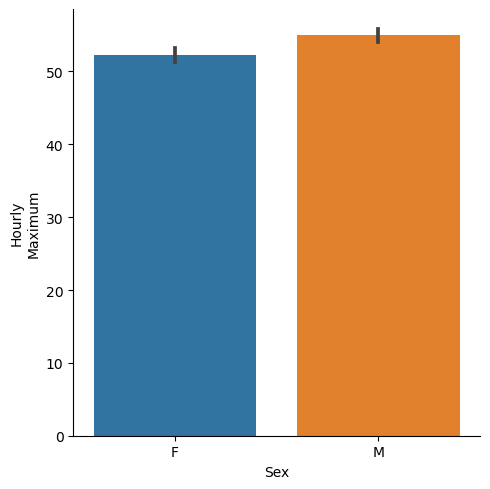
After this I used a NumPy function to determine the hourly min of the employees by and organized by their race and gender. I moved to visualization and decided Seaborn would best suit the dataset I had used (Data Camp, *Introduction to Seaborn*, 2023).

# **Results**

Chart, bar chart

Description automatically generated

The first chart I created looked to visualize the minimum hourly salary of the employees, which I plotted on the Y axis, and looked to the employees race, which I plotted on the X axis. I chose to use a catplot chart as the wick on the top of the bar showed the variation of minimum salary for each race.



The next chart I chose to run was another catplot bar chart, but his time I wanted to see was there a difference between the genders and the maximum hourly salary an employee could receive.

Chart

Description automatically generated

The final chart I ran was a catplot bar chart, I wanted to display clearly the breakdown of each race of the employees in Arlington county. All of the charts I produced I wanted to be easy to read and clear, as in my work I often have to present findings and charts and often having an easy to read chart proves more useful when presenting findings to senior staff.

# **Insights**

* The average start date for men is a year earlier than women, which may account for some of the wage differential between genders.
* White and Asian employees are highly paid but there is far more variation in Asian employees hourly pay.
* On the lower end of the wage scale Black/African American and Hispanic/Latino are paid less, however there is little variation between these employees hourly minimum salary.
* When comparing female and male hourly maximum pay, men are paid more with similar levels of variation in both genders.
* While the largest minority groups are black/African American and Hispanic/Latino employees they are among the lower paid employees within Arlington County. Interestingly, the Asian employee population which is small are among the highest paid employees.
* The only racial group in which women are paid a higher average hourly minimum is Native Hawaiian or Other Pacific. Women in this group are paid $8.25 more than their male counterparts.

**Machine Learning:**

* **Describe what kind of prediction you could perform in the future using machine learning and/or deep learning?**

A prediction that could be made using machine learning or deep learning on the data, or similar data would be as a predictor for whether staff would need pay raises in the coming future to coincide with inflation. If inflation and cost of living data was input via unsupervised learning, it could predict a down turn in the Arlington county economy that could signal to the county officials to put a hold on recruitment for a predicted amount of time (Sotayo, Slide 12, 2023).

Supervised learning would be best suited as it is labelled. As this data looks at demographic data most predictions would involve groups of people, whether that be by gender or race, using the regression method would work best in obtaining predictions (Data Camp, *Supervised Learning*, 2023).

* **Would you use classification or regression methods?**

In respect to the dataset I examined during this project. The appropriate statistical model I would apply is that of regression (Sotayo, A, Slide 10, 2023). This could be applied to predict the trajectory of the state employees pay scale. Whether it is likely to increase or decrease. The regression model would suit best as it assigns a continuous variable, which is likely to be the result of a model applied to this dataset (Data Camp, *Machine learning concepts*, 2023). It would be interesting to also pull data from cost of living and inflation in the Arlington county to see if the pay scale of the employees was to coincide with rising cost of living.

It would also be interesting to run a regression model to see if the pay differences between genders and particular races is likely to increase or decrease in the future. It would be interesting to pull demographic and population data from a census of Arlington county. This could allow for a deeper insight to inequalities that are present for county employee.

# References

campus.datacamp.com. (n.d.). *Machine learning concepts | Theory*. [online] Available at: <https://campus.datacamp.com/courses/understanding-machine-learning/what-is-machine-learning?ex=5> [Accessed 24 Jan. 2023].

‌

campus.datacamp.com. (n.d.). *Supervised learning | Theory*. [online] Available at: <https://campus.datacamp.com/courses/understanding-machine-learning/machine-learning-models?ex=1> [Accessed 24 Jan. 2023].

‌

campus.datacamp.com. (n.d.). *Introduction to Seaborn | Python*. [online] Available at: <https://campus.datacamp.com/courses/introduction-to-data-visualization-with-seaborn/introduction-to-seaborn?ex=1> [Accessed 24 Jan. 2023].

‌

Data.gov. (2020). *Employee Demographics*. [online] Available at: <https://catalog.data.gov/dataset/employee-demographics>. [Accessed 21 Jan. 2023].

‌

Fiscal Year 2023 Pay Schedule For County Employees. (2022). [online] Available at: <https://www.arlingtonva.us/files/05ae152f-fad5-436d-b90c-7ebde30973a7/FY23A-Pay-Plan-for-County-Employees.pdf> [Accessed 21 Jan. 2023].

‌

Sotayo, A. 2023. Machine Learning Lecture Slides. <https://learn.ucdpa.ie/pluginfile.php/90104/mod_resource/content/3/Unit8-DataAnalyticsCertificate-MachineLearning.pdf> [Accessed 23 Jan. 2023].