Growth of the Airline Industry

Erya Ma

Wendy Huang

Gloria Hu

Summary of Research Questions

1. **What is the trend of revenue passenger miles of each type of operation? Which year has the highest RPM? Which year has the lowest RPM? What's the trend of overall RPM?**
   a. Revenue passenger miles are calculated by multiplying the number of paying passengers by the distance traveled. We expect the revenue of passenger miles to increase over the years. To visualize the trend, we will plot a multi-line chart. We also want to know the trend of overall RPM, and which year has the highest and lowest RPM. We found out that 2015 has the highest RPM and 2003 has the lowest RPM, which is slightly different from what we expected. We expected 1995 to have the lowest RPM. However, the overall trend is as we expected.

2. **Which type of aircraft(small narrowbodies, large narrowbodies, and widebodies) has the highest salary for a pilot over the years? Do they have similar salary trends over years?**
   a. To compare the salary of different aircrafts and explore the salary trends over years, we will filter the salary and wages of each aircrafts in each year by using the data from UA_aircraft_operating_stat.csv first, then plot a line graph to visualize the salary trend. We predict that pilots' salaries have a positive correlation with the size of the aircraft, however, we found that large narrowbodies have a lower salary than the small narrowbodies.

3. **What's the trend of overall UA Average salaries for each type of employee over time? Is there any correlation between the average salary s and each type of**

employee position? Explain what you've found through each diagram and point out important insights.

a. To explore if UA average salary has any correlation with each employee position, we will be using UA_salary_benefits.csv data set to plot five diagrams: Average Salary for Management & All Other, Pilots & Co-Pilots, Flight Attendants, Maintenance, and Passenger, Cargo & Aircraft Handling. We expect that most employee's type/position should have a positive correlation with the salary they gain every year.

Motivation and Background

The motivation behind choosing this topic is to see the growth of United airlines over the years. Our initial thought is to reveal the dramatic changes to airline industry revenues and expenses. Traveling by air has always been the most expensive type of transportation, but with the development of the economy, people have become more and more wealthy. Traveling by air has become a very common type of transportation. As more and more people travel by air, the revenue has increased dramatically. In the meantime, the operating expenses, such as labor, fuel, maintenance and the outsourcing of some services and departments, have also increased dramatically. We want to explore the potential correlation or related association between elements and use Python to illustrate the impact or valuable findings better.

Dataset

Link: https://data.world/adamhelsinger/united-airlines-data

Summary: The US commercial airline business is one of the most diverse, dynamic, and complex in the world. It is fast-evolving, labor-intensive, capital-intensive, hyper-competitive, and extremely subject to business cycle ebbs and flows as well as one of the most regulated of deregulated industries. The United airlines data was collected by the MIT Global Airline Industry Program and stored in four excel data tables. Data is about traffic, capacity, operation, aircraft operation, and employee data of united airlines.

Methodology

For question one, we will reformat the traffic and capacity data frame, and create a new csv named operation summary. The first column includes 5 types of operations which are Atlantic operations, latin america operations, pacific operations, domestic operations, and international operations. We will plot a multiple line graph using seaborn with year in axis and Sum of RPM (Revenue Passenger Mile) in y axis. We will also return a dictionary with the year as the key and the sum of revenue passenger miles as the value. Then we can find the year with the highest RPM and lowest RPM.

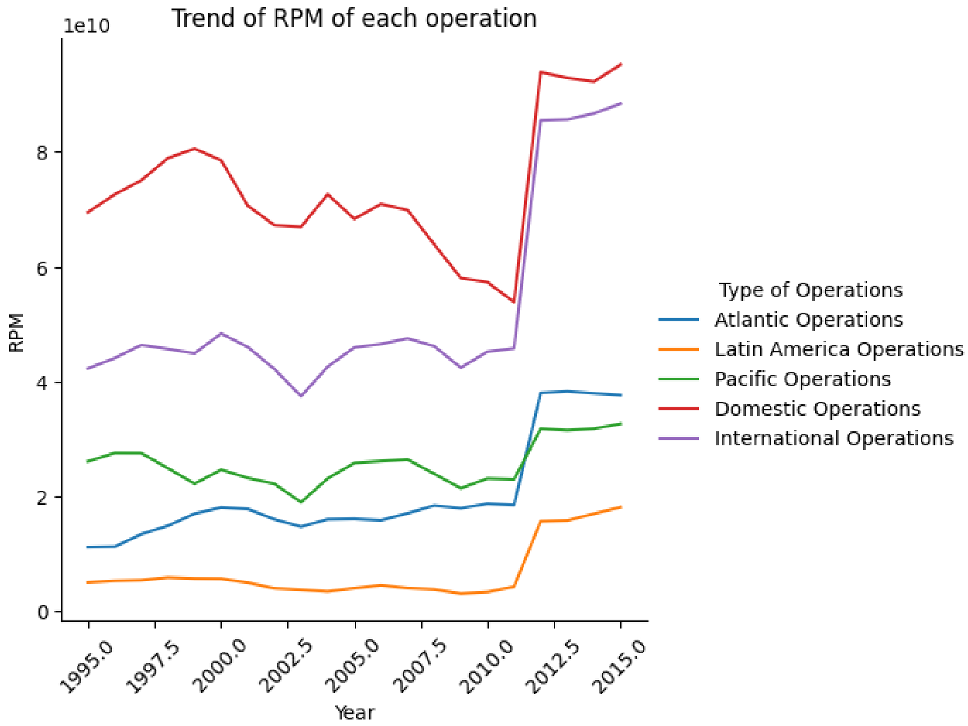
For question two, we will filter the United Airlines Aircraft Operating Statistics data frame to only three columns: aircraft type (including small narrowbodies, large narrowbodies, widebodies, and fleet), salary and wages, and years, so that we can grouby the aircraft type and find the maximum salary over year. Finally we can use the Matplotlib library to put a line graph for our result, which will allow each line to represent each aircraft type.

For question three, we will first recognize columns for each job position and Salaries/Wages. Since each position is separately as titles, we need to filter and reformat the data frame. By using the convert method, we are able to use data as floats instead of string in the organization process. We will use the droppna method to get rid of the variables that don't have values in our set. In our plotting section, we will plot each employee position with their average salary over time by using regression plot, dot plot, and histogram to explore potential association or correlation between average salary and employee job position.

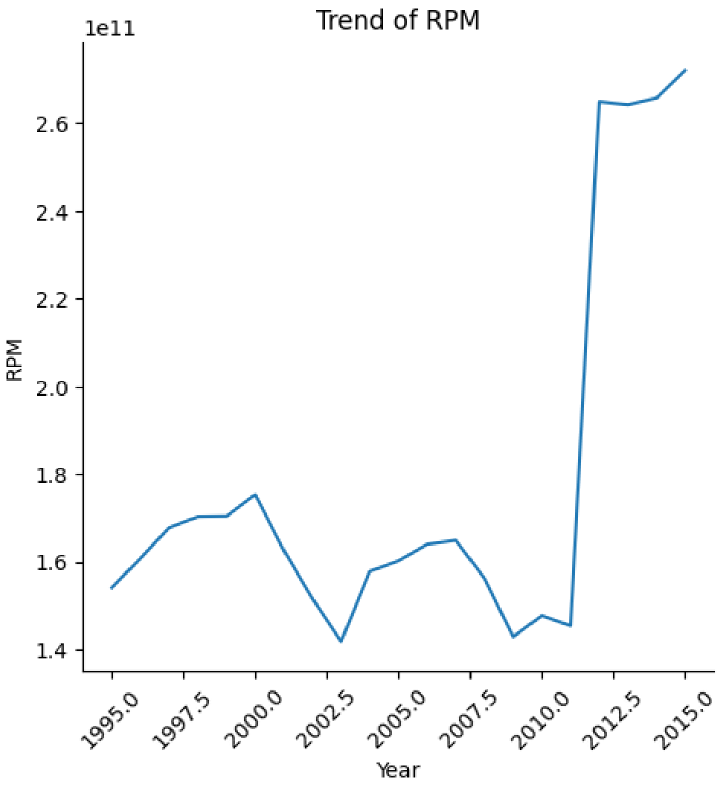
Results

1. **What is the trend of revenue passenger miles of each type of operation? Which year has the highest RPM? Which year has the lowest RPM? What's the trend of overall RPM?**
    a. A revenue passenger mile (RPM) is a transportation industry metric that shows the number of miles traveled by paying passengers and is typically an airline traffic statistic.

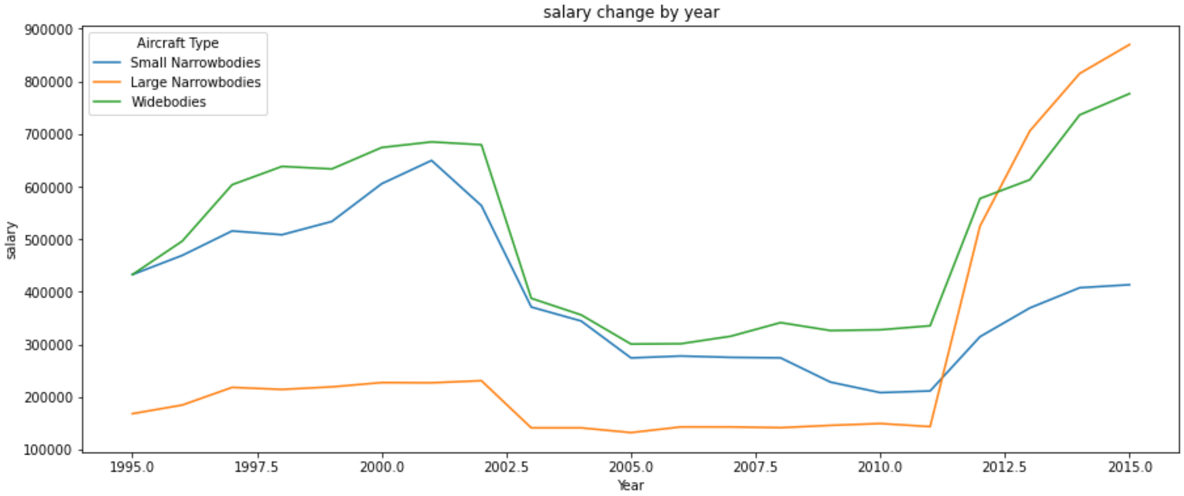
By looking at the graph ‘trend of RPM of each operation’, the domestic operation has the highest average RPM, and the Latin America operation has the lowest average RPM. One similarity for all operations is that RPM increased dramatically in 2011.



b. By looking at the graph ‘Trend of RPM’, RPM is the lowest in 2003 and highest in 2015. The overall trend of RPM is increasing but not very stable with a dramatic increase in 2011.



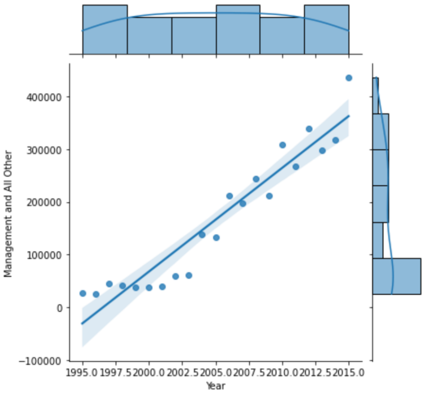
2. **Which type of aircraft(small narrowbodies, large narrowbodies, widebodies, fleet) has the highest salary for a pilot over the years? Do they have a similar salary trend over time?**

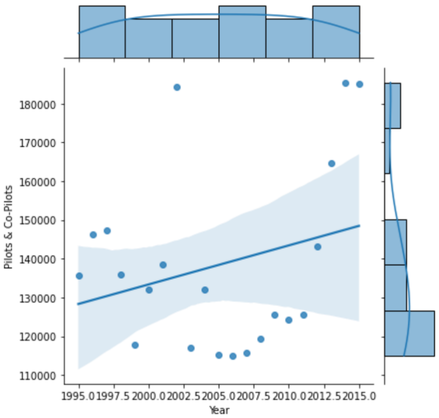


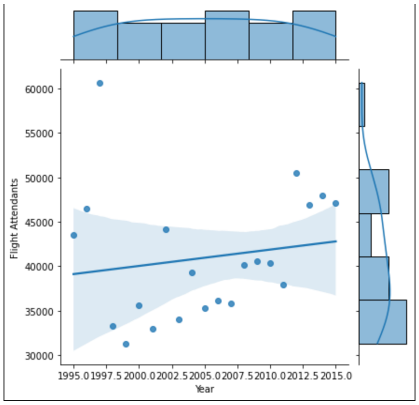
[Year vs. Salary of pilots ($)]

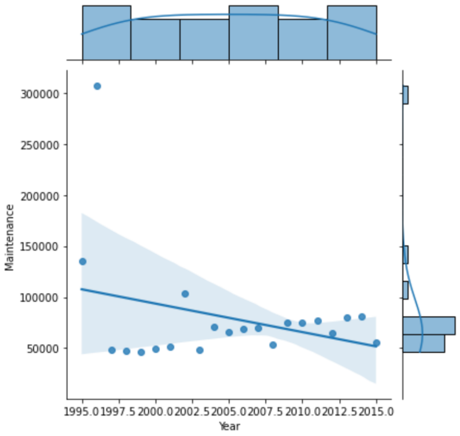
The result in this graph is very interesting and unexpected. We expect that larger the airplane is, higher salary the pilots have. However, we saw that widebodies provide the highest salary before 2011, and large narrowbodies have the lowest salary. And the widebodies and large narrowbodies have a similar salary trend while the widebodies have an extremely low and different salary trend. We suspect that there was a sudden salary drop in 2001 with the huge shock due to the 911 attack, the airline industry lost millions of profits as people felt insecure traveling with planes, and the industry needed to put more budgets on security updates. We also notice that there was a great salary increase in 2011 for all types of aircraft pilots, which suggests that something changed around that time, either a policy that benefits the industry or the economic recovery from the financial crisis. What's more, the positive trend in salary after 2011 also represents the rapid development of the economy in the US. And it's surprising that after 2011, pilots who work in the large narrow bodies have higher salaries than the other two aircraft, while they have the lowest salary before 2011.

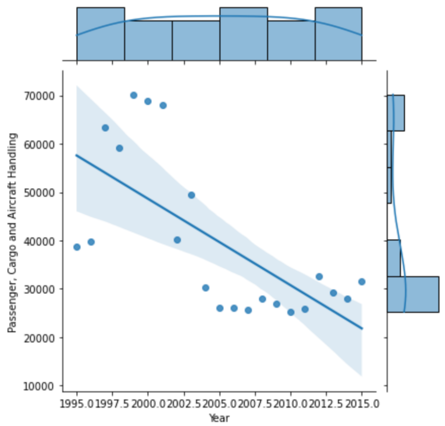
3. **What's the trend of overall UA Average salaries for each type of employee over time? Is there any correlation between the average salary and each type of employee position? Explain what you've found through each diagram and point out important insights.**

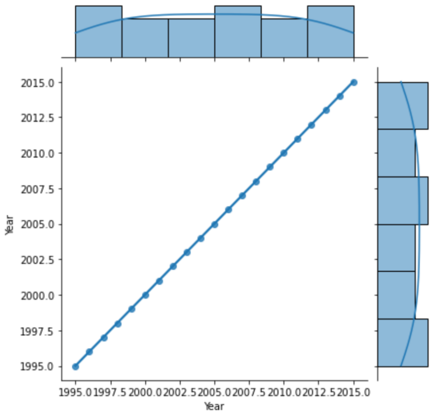


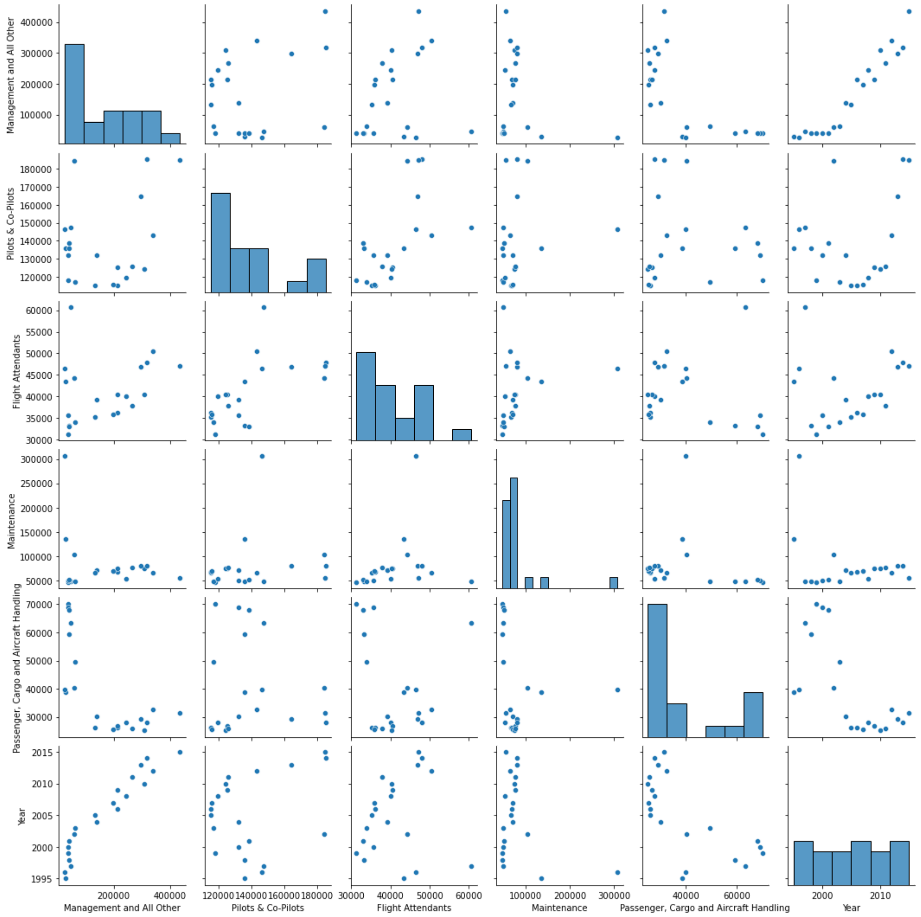












Our data visualizations show that the average salary for management positions increases solidly every year, and there is a strong association between those two from the graph. As for pilots and co-pilots, their average salary increased at a steady rate from 1995 to 2015, and there was a positive correlation between pilots/co-pilots and their average salary. Compared to pilots, flight attendants had a lower increase rate in average wage every year. As time went on, employees for work for maintenance had a moderate decline in average salary. Employees who take care of passenger, cargo, and aircraft handling face a tremendously decreasing average wage. More advanced technology adopted in airports, such as automatic conveyor belts with a faster function, will replace part of this type of employee's work. There is a strong negative correlation between employees caring for passengers, cargo, and aircraft handling and their average salary. The overall plot demonstrates the highest average salary paid employees are working for management or other services at UA. Pilots and flight attendants have an increasing salary every year. The lowest average salary paid are employees who work handcrafting.

Impact and Limitations

Based on our results, we analyzed our direct and indirect stakeholders. UA airline companies can make better decisions based on airline spendings, operation summary, and employee's average salaries. Companies will have a better control of their costs in multiple areas.

As we examined from our dataset, employees who work at management services had the highest average salary paid over time. One limitation in this dataset is that it has a vague definition of the position title "Management and All Others". We don't know exactly what other types of employees are part of this category. As for new audiences or readers, it may cause confusion and misunderstanding. As for some employee's having a decreasing rate in averay salary, which will result in discouraging employees in their careers.

Since the data we use is all from UA, we can not use the result to make inferences about the entire industry. We can assume that UA's data is valuable to the industry and industry-wide trends are similar to those in UA because UA is a big company. But we can not conclude that this is the case. Each country's situation is different, so our analysis has no reference value for the air transportation industry of other countries.

Challenge Goals

1. We will use **multiple datasets** in our project, and **data is messy**. We need a format that would be useful for the analysis. We need multiple datasets to come up with a richer analysis. And some questions require us to join tables together to solve. We will take time

to consider which data should include in our dataframe and how to reformat it. Through collaborative effort, we can achieve our goal.

2. We will use **various methods** in our project to make our data visualization **insightful** and **clear** to understand. Some of the datasets need to be sorted and **re-filtered** by **creating** new columns, **converting strings to float**, **dropping** unnecessary variables, and reorganizing the required dataset. Our plot formats are various: **line chart, multiple-line diagram, histogram, dotplot**, and we implement a **regression line** to our visualization for more significant analysis. We will use what we've learned to make a solid and concise data visualization to promote both **input and output**.

Working Plan Analysis

1. **Data Cleaning | Predicted: 4 hours | Actual: 6 hours**
    a. All four datasets are very messy and are not suitable for python analysis. So we manually reformat the dataset to a more useful format. We took longer than expected to figure out what data we needed and how to reformat.
        i. Our group collaboratively converted the dataset to a format that would be useful for the analysis.
            1. Gloria: Focus on reformatting datasets
            2. Lyra: Handle data that are missing
            3. Wendy: Debugging and supporting code examples from previous class materials to refer.
2. **Data Manipulation | Predicted: 8 Hours | Actual: 9 hours**
    a. Question 1 (Lyra): add up revenue passenger miles of each operation, rank revenue passenger miles in descending order, then find the year with the largest revenue passenger miles. We test our code using a different way to calculate the same result. with
    b. Question 2 (Wendy): create a new dataframe. Filter dataset by the type of aircraft and groupby with salaries and wages. We test our code using a debugger and common sense.
    c. Question 3 (Gloria): create a new dataframe. Filter dataset by average salary and each employee type from 1995 to 2015. We need to convert values from string to float to make them usable.

3. **Plotting | Predicted: 4 Hours | Actual 4 hours**
    - **a.** Question 1 (Lyra): plot a line chart with years on x-axis and sum of revenue passenger miles on y-axis. Plot a multi line chart with years on x-axis and total revenue passenger miles on y-axis.
    - b. Question 2 (Wendy): plot a multiple line graph with years on x-axis and salaries and wages on y-axis. Different types of aircraft are distinguished by color.
    - c. Question 3 (Gloria): plot multiple graphs to visualize the relationship or correlation between each employee type and their average salary over time. Use a paired-plot method to list all diagrams for data visualization analysis
4. **Report Writing | Predicted: 10 Hours | Actual: 10 hours**
    - a. Each member writes one or two sections of the report.
        - i. Analyze if there is any correlation or association from the data we are analyzing. And demonstrate if there is any limitation of this dataset – Gloria
        - ii. List the findings and explain our prediction to future growth trend of U.S. Airlines – Gloria
        - iii. Demonstrate our report with the supportive evidence from our plots. – Lyra

Testing

In general, we tested our code using debugger and common sense. To test the result of the first question, we intend to find the sum of RPM of all operations in each year without using pandas. If the result which is calculated without pandas is the same with the result of rpm_year_sum, then we can ensure your result is correct. To test the result of the year with highest RPM and lowest RPM, we compared each value in rpm_year_sum to identify which year is the highest and which year is the lowest. We used assert statements to do so. Everything we saw made sense, so we passed it.

Collaboration

We did not collaborate with people outside the class