Mod 8, Opt 1: Capstone Project: US Organization

Alicia Unke

Colorado State University Global

MIS581 Capstone - Business Intelligence and Data Analytics

Osama A. Morad, Ph.D.

Nov. 29, 2020

Alaska Airlines Flight Delays Project

**Abstract**

In order for Alaska Airlines to maintain their reputation for providing the best customer service it is important that they keep their customers happy by providing flights that arrive on time. By monitoring flight data, Alaska Airlines can come up with real time solutions that will prevent huge monetary losses from delayed flights.

**Background**

Alaska Airlines was founded over 85 years ago when Barnhill & McGee Airways merged with Star Air Service. “After several more mergers, where many more carriers were woven into the fabric of the company, the Star Air Service name was changed a couple of times - until 1944, when they found one that stuck: Alaska Airlines, the name we use today” (2020). By 1972, Alaska Airlines found themselves to be in a financial crisis and two business men set out to start winning over the trust of creditors. They came up with a new business plan to set themselves apart from the other airlines and established ways to improve on-time flight performances. How they were going to set themselves apart from other industries was by providing outstanding customer service. This philosophy worked so well that it helped the company grow and allowed for nineteen straight years of profitability (2020). Over the next decade, Alaska Airlines had tripled their fleet in size and expanded by joining forces with Horizon Air and Jet America.

Today, Alaska Airlines continues to put their customers first and continue to stand out for their great customer service. They continue to grow by offering more flights to new destinations such as the east coast and to Hawaii. Alaska Airlines is currently headquartered in SeaTac, WA and they have over 23,300 employees working nationwide. In 2018, their revenue was around $8.264 billion with a net income of $437 million (Wikipedia, 2020). As of 2017, they maintain a fleet of 248 jet aircrafts and their subsidiary airline, Horizon Air, has a fleet of 41 turboprop and 16 jet aircrafts. In total Alaska Airlines has a combined fleet of 298 aircrafts.

**Dataset**

The dataset that will be used in this data analysts research project is from the U.S. Department of Transportation website. The dataset is managed by the Bureau of Transportation Statistics department and contains data involving all the major airlines flight information. Since the database is so large, I will be narrowing down the dataset to just using Alaska Airlines flight data from the last 2 years (Aug. 2018 to Aug. 2020). The purpose of the dataset is to track whether flights are arriving on time or if they are delayed. Any flight arriving 15 minutes past the scheduled time is considered delayed and affects Alaska Airlines performance. In total, this dataset has 16 variables including two variables for the month and year of the flight.

According to the U.S. Department of Transportation website, the definition of each flight delay category are as follows:

* **Air Carrier:** The cause of the cancellation or delay was due to circumstances within the airline's control.
* **Extreme Weather:** Significant weather conditions (actual or forecasted) such as tornado, blizzard or hurricane.
* **National Aviation System (NAS):** Delays and cancellations attributable to the national aviation system, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control.
* **Late-arriving aircraft:** A previous flight with the same aircraft arrived late, causing the present flight to depart late.
* **Security:** Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas. (US DOT, 2020).

**Data Dictionary**

Figure 1 below shows a data dictionary of the Alaska Airlines flight dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Field Name | Data Type | Data Format | Field Size | Description | Example |
| year | Date | YYYY | 4 | Year of flight date | 2017 |
| month | Date | MM | 2 | Month of flight date | 06 |
| carrier | Text | XX | 2 | Carrier abbreviation | AS |
| carrier\_name | Text |  | 20 | Carrier name | Alaska Airlines |
| airport | Text | XXX | 3 | Airport abbreviation | SEA |
| airport\_name | Text |  | 50 | Airport city, state and name | Seattle, WA: Seattle/Tacoma International |
| arr\_flights | Integer | xxxxx | 5 | # of flights arrived on time | 4551 |
| arr\_del15 | Integer | xxxxx | 5 | # of flights arrived 15 min after scheduled arrival time | 1045 |
| arr\_cancelled | Integer | xxxxx | 5 | # of flights cancelled | 75 |
| arr\_diverted | Integer | xxxxx | 5 | # of flights diverted | 7 |
| arr\_delay | Integer | xxxxx | 5 | Total # of flights delayed | 57228 |
| carrier\_delay | Integer | xxxxx | 5 | # of flights delayed due to circumstances within the airline's control | 10085 |
| weather\_delay | Integer | xxxxx | 5 | # of flights delayed due to bad weather | 3207 |
| nas\_delay | Integer | xxxxx | 5 | # of flights delayed due to the national aviation system | 20421 |
| security\_delay | Integer | xxxxx | 5 | # of flights delayed due to security issue | 331 |
| late\_aircraft\_delay | Integer | xxxxx | 5 | # of flights delayed due to previous delayed flight with same aircraft | 23184 |
|  |  |  |  |  |  |

Figure 1: Alaska Airlines Flight Dataset Data Dictionary

The biggest reason I chose a public dataset that tracks airline delays is because airline delays greatly affect profits and revenues. Airlines that have flights that are consistently delayed risk losing customers which in turns causes the company to lose profits.

**Hypothesis**

One of Alaska Airlines main goals is to provide the best customer experience. One way to do that is to ensure flights arrive and take off on time. For this portfolio project, the business questions we will be trying to answer along with the hypothesis are:

* What reason contributes to the most number of flight delays?

Null Hypothesis: The most flight delays are because of the Air Carrier.

Alternate Hypothesis: The most flight delays are not because of the Air Carrier.

* Is there a greater number of flights arriving on time vs the flights that are delayed?

Null Hypothesis: There is not a number percentage of on time flights then delayed flights.

Alternate Hypothesis: There is a greater number of on time flights then delayed flights.

**Tools and Techniques**

Christina Majaski (2020) states that there are four steps for hypothesis testing and my portfolio project will follow these steps. The four steps are:

1. State the two hypotheses so that only one can be right.
2. Formulate an analysis plan which outlines how the data will be evaluated.
3. Carry out the plan and physically analyze the sample data.
4. Analyze the results and given the data, either reject the null hypothesis or state that the null hypothesis is plausible (Majaski, 2020).

Analysis of the dataset followed the extract, transform, and load (ETL) process to ensure accurate results. Once the statistical models had been runned, graphs and visualizations were made of the results.

When analyzing the Alaska Airlines dataset, the summary statistics was performed by using SAS Studio and the Sum and Mean command was used in Python Jupyter. Visualizations were then created in Python Jupyter. The results were analyzed to determine if the null hypothesis should be rejected or not.

**Security, Privacy, and Ethics**

Hackers are always trying to access companies databases and gain private information that they can sell. It is important for companies to protect their customers' data as best as possible. Big datasets have many security concerns due to their size. Some security issues are caused by: cloud configuration problems, weak identity governance, clashing application softwares, glitchy new software, and remote data storage (Rice, 2019).

“Ethical challenges occur when opinions on what is considered right and wrong deviate” (Jha, 2019). Unfair discrimination, reinforcement of human biases, and lack of transparency are three ethical challenges that relate to data analysis. Unfair discrimination involves data that only relects a certain group of people. For example, only using data from a certain race or gender and excluding the other ones. Reinforcement of human bias occurs when computer models are used to make predictions. Artificial Intelligence (AI) is a trained model and when a certain group has been known to contribute to something the model might deem everyone in the group the same even if it doesn't apply to them. Lack of transparency involves “the step-by-step process, model and its parameters by which a prediction is made and it remains unclear which data is being used in making a prediction” (Jha, 2019). There are not many laws regulating data ethics, so one challenge is defining what is considered ethical and not while pulling data for a project.

**Data Analysis Results**

In order to find the answer to the first hypothesis question, which is ‘What reason contributes to the most number of flight delays?’, a summary statistics test was performed. The “proc means” command was used in SAS Studios to find the results. Figure 1 below shows the results broken down by the year and Figure 2 shows the similar results but instead of being broken down by the year, it is the averages for the whole dataset.

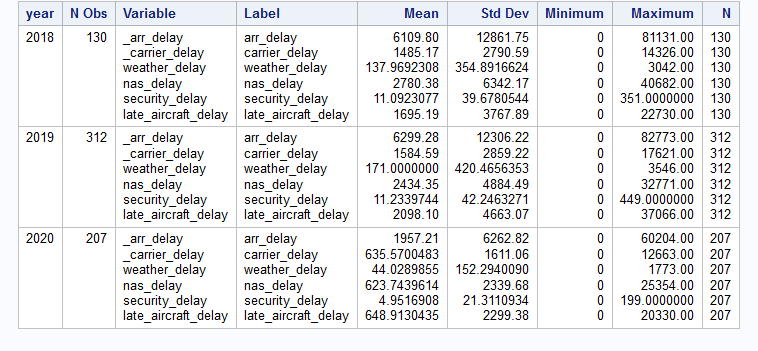


Figure 1. SAS Studio. Average number of airlines delays per year. Screenshot taken by author.

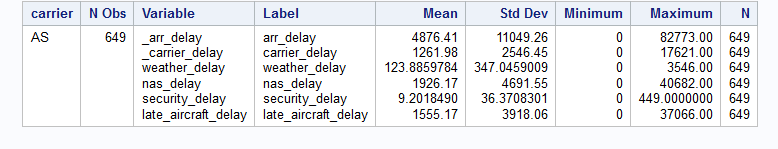


Figure 2. SAS Studio. Average total number of delayed flights. Screenshot taken by author.

In Figure 1 and Figure 2 above, “\_arr\_delay” refers to the total number of airlines that were delayed. The next five categories are the reasons why the flights were delayed. Looking at the two charts it looks like on average, the biggest reason for delays was because of NAS. This refers to the overall airport operations such as air traffic control, air field congestion, or unexpected bad weather. The second biggest reason for delays is because of late aircraft delays. This means if an incoming flight arrives late then the following flight is also considered late because it has to leave later than the expected departure time. This is important information for airlines because schedules need to accommodate for late arrivals in order to stop the chain of events of late departures from happening. While the Alaska Airlines can’t control delays due to NAS, they can control late aircraft delays. Based on this analysis my alternate hypothesis was correct.

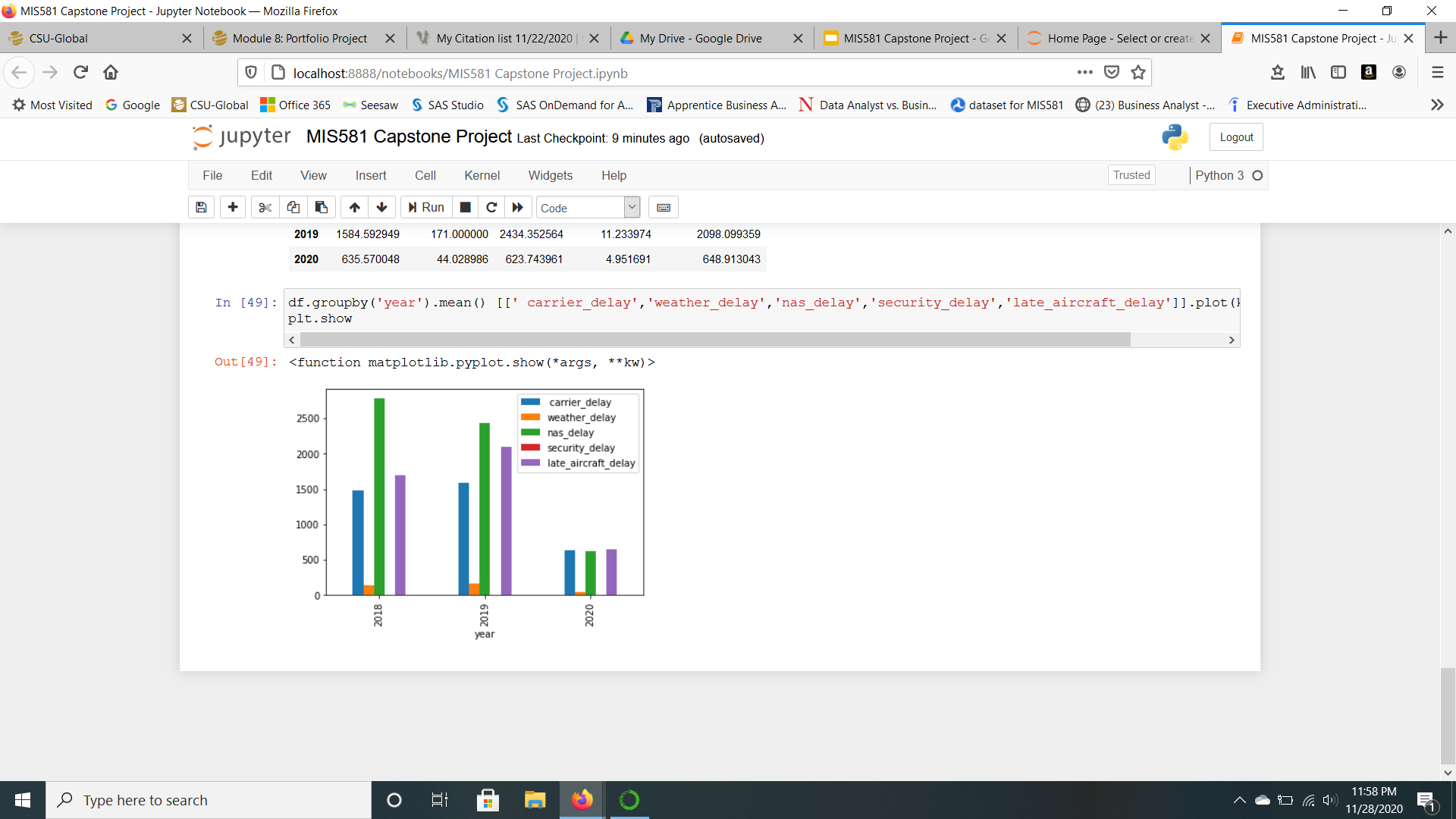
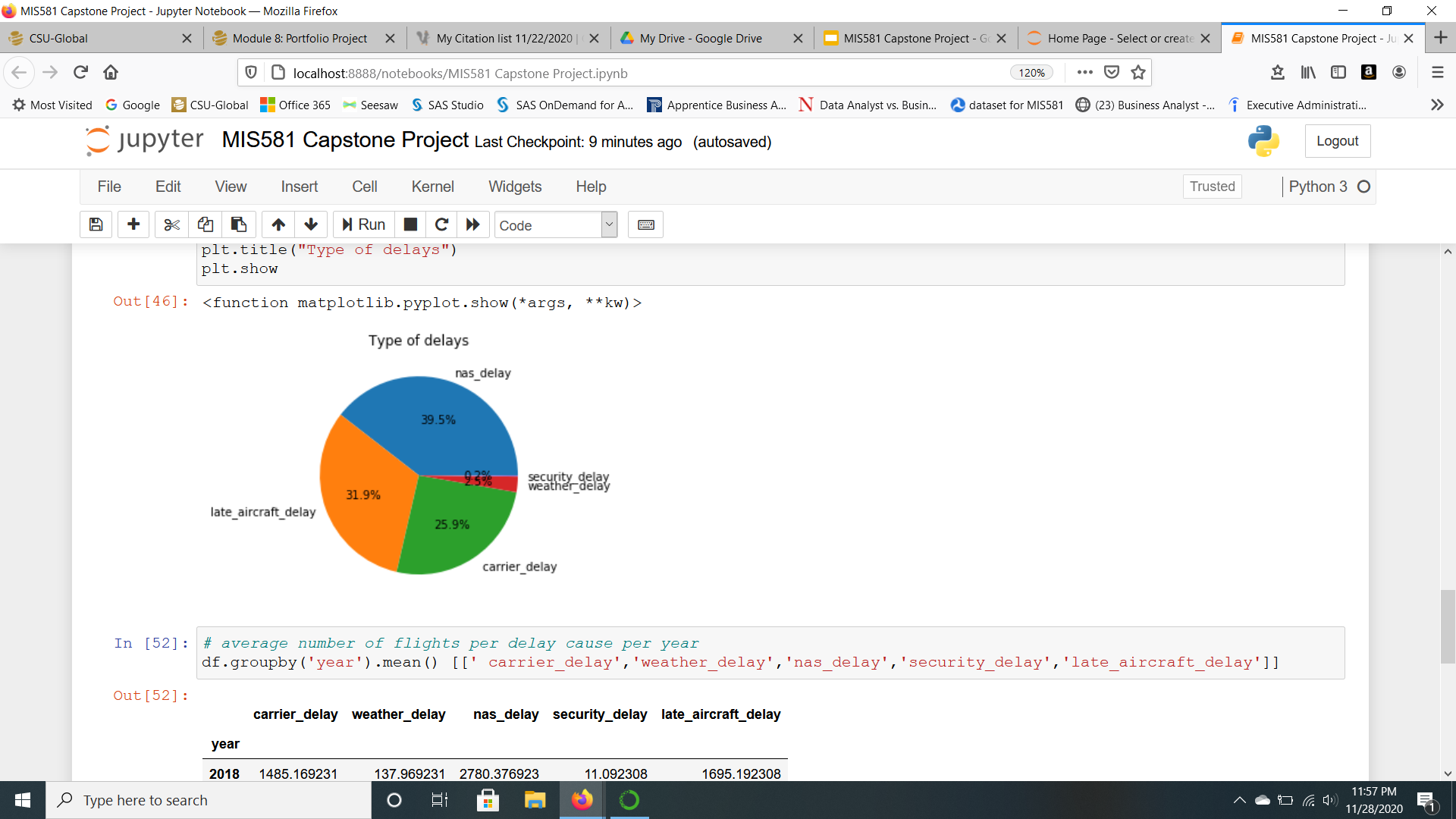
 

Figure 3. Python Jupyter. Visualization of results. Screenshot taken by author.

Figure 3 provides a visualization of the data results shown in Figure 1 and Figure 2. The bar graph on the right shows the results for reasons why flights are delayed per year. The pie chart on the right shows the overall percentages of each flight delayed group. Both graphs help us determine that we need to reject the null hypothesis for the first business question.

Figure 4 table below shows the results of the analysis for answering the second hypothesis question. The second hypothesis question is ‘Is there a greater number of flights arriving on time vs the flights that are delayed?’ and based on the results on average it looks like most flights arrived on time. This means my alternate hypothesis was correct for this question as well and that we need to reject the null hypothesis.

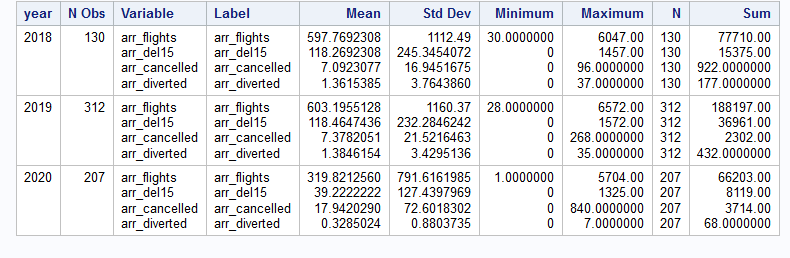


Figure 4. SAS Studio. Number of flights per category per year. Screenshot taken by author.

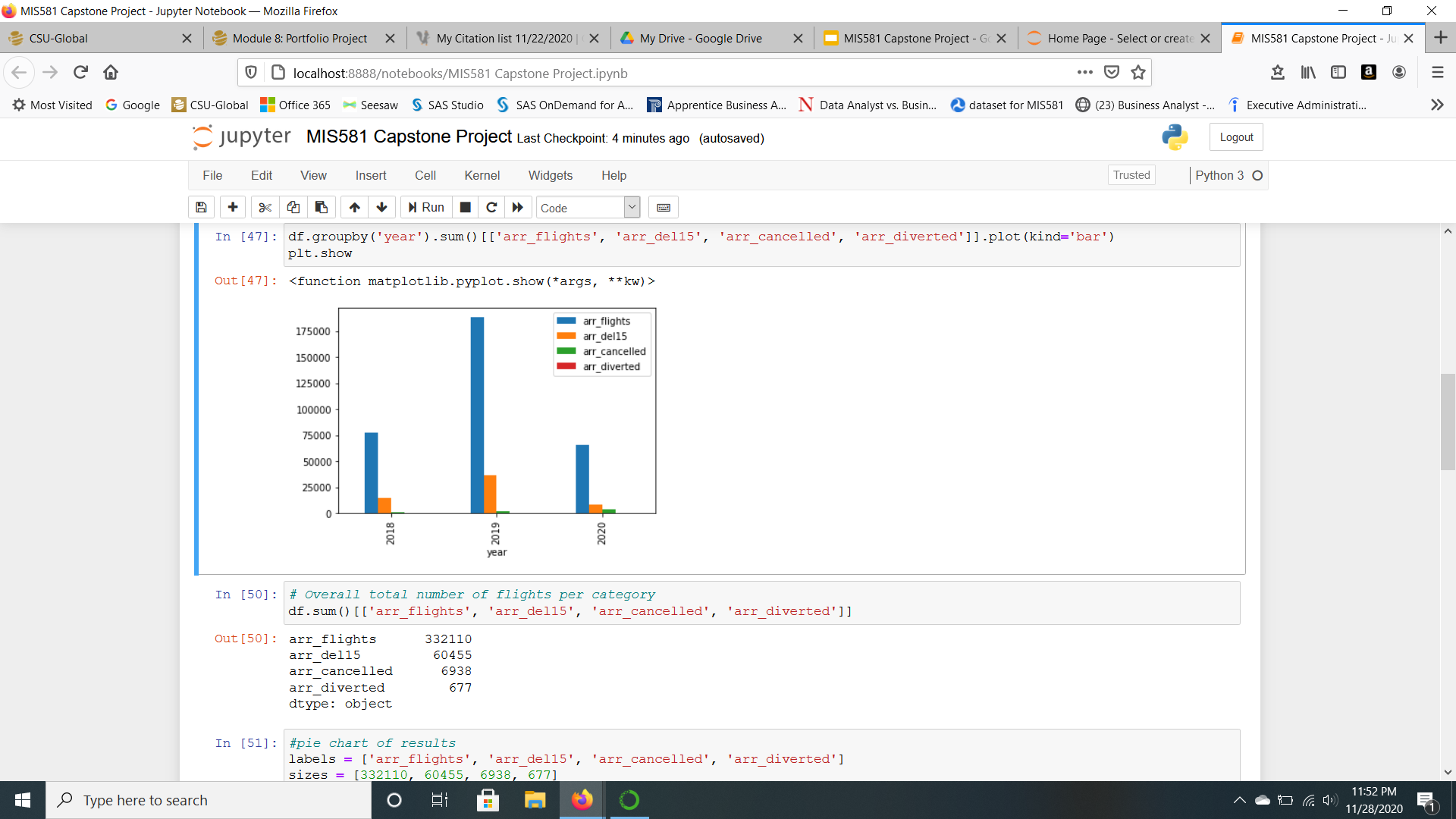
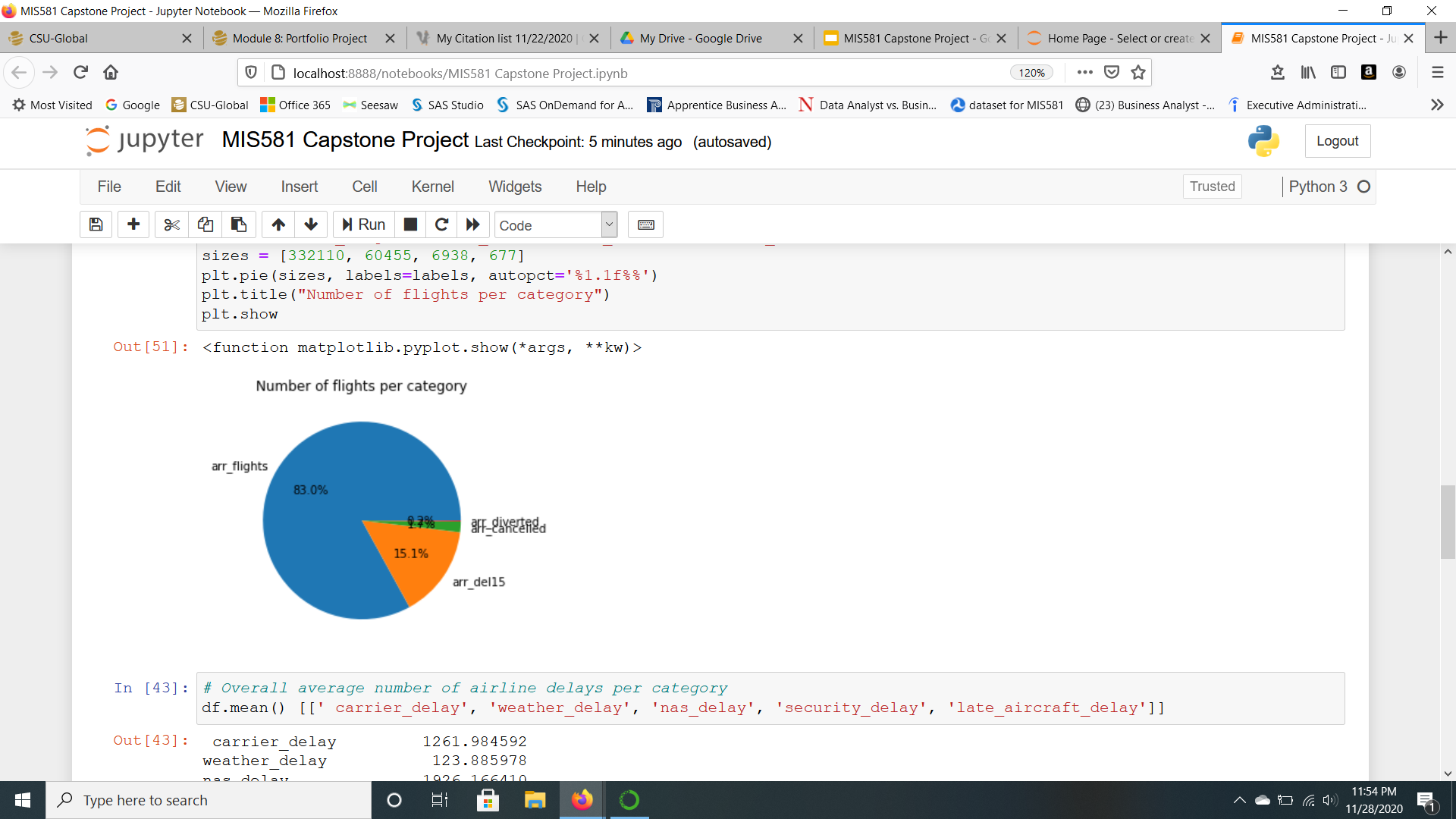
 

Figure 5. Python Jupyter. Visualization of results. Screenshot taken by author.

Figure 5, above, displays the results for the second business question in two different graphs. The bar graph shows the data outcome per year and the pie chart shows the overall percentage of each category. The graphs are based off of the “Sum” results in Figure 4’s table. Looking at the graphs we can easily see Alaska Airlines flights arrive on time way more than being delayed, cancelled, or diverted.

**Programming Code Used**

All programming code for this project was uploaded to my GitHub account at the following link: <https://github.com/aunke/AlaskaAirlinesFlightDelays.git> . In SAS Studios, “proc means data” command was used to find the summary statistics of the data. In Python Jupyter, “sum()” and “mean()” commands were used to provide additional analysis of the data set. Pie charts and bar graphs were also created using matplotlib and these commands provided visuals for the project.

**Conclusion**

Alaska Airlines can learn a lot from the analysis and data set because preventing delays helps the company maintain satisfied customers. Since the results show that late aircrafts are the second biggest reason on why aircrafts are delayed, Alaska can take this information and adjust their flight schedules to prevent this from occurring. By decreasing delays and improving profits, Alaska Airlines will continue to be able to invest in new innovations and provide the best customer service.

**Literature Review**

Alaska Airlines History. (2020). Retrieved from <https://www.alaskaair.com/content/about-us/history>

This website is very helpful in this research paper because it provides us with a background on Alaska Airlines. There is lots of information about Alaska Airlines history and business plan. Research Alaska Airlines history provides us with our information for the introduction.

Jha, S. (2019, January 04). The Ethical Challenges Of A Data Science Practitioner. Retrieved November 01, 2020, from https://blog.quantela.com/the-ethical-challenges-of-a-data-science-practitioner/

While researching ethical challenges, this article popped up. It has a lot of useful information regarding ethics and was used in the ethics section of this paper to provide evidence of ethical challenges.

Majaski, C. (2020, January 27). Hypothesis Testing. Retrieved from <https://www.investopedia.com/terms/h/hypothesistesting.asp>

This resource provides valuable information on hypothesis testing. It also provides helpful steps on how to properly test a hypothesis question. This article was also used as a guideline on how to create a strong hypothesis question.

Rice, M. (2019, August 26). 7 Big Data Security Issues to Consider. Retrieved November 28, 2020, from <https://builtin.com/big-data/big-data-security>

This resource was used to find out more information on database privacy and security. It contained a lot of good information regarding different security issues.

US DOT. (2020, March 05). Airline On-Time Performance and Causes of Flight Delays. Retrieved from <https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays>

The U.S. Department of Transportation website is where the dataset used in this analysis is from. This website also provides us with important information on what each variable in the dataset means. There are five different reasons why flights are delayed and this source gives us examples for each reason.

Wikipedia. (2020, October 05). Alaska Air Group. Retrieved October 11, 2020, from <https://en.wikipedia.org/wiki/Alaska_Air_Group>

Wikipedia was used for finding out more information about Alaska Airlines. The article had useful information such as: the company’s aircraft fleet size, employee size, annual revenue, and where it is headquartered.

References

Alaska Airlines History. (2020). Retrieved October 11, 2020, from <https://www.alaskaair.com/content/about-us/history>

Jha, S. (2019, January 04). The Ethical Challenges Of A Data Science Practitioner. Retrieved November 01, 2020, from https://blog.quantela.com/the-ethical-challenges-of-a-data-science-practitioner/

Majaski, C. (2020, January 27). Hypothesis Testing. Retrieved October 22, 2020, from <https://www.investopedia.com/terms/h/hypothesistesting.asp>

Rice, M. (2019, August 26). 7 Big Data Security Issues to Consider. Retrieved November 28, 2020, from https://builtin.com/big-data/big-data-security

US DOT. (2020, March 05). Airline On-Time Performance and Causes of Flight Delays. Retrieved October 11, 2020, from https://www.bts.dot.gov/explore-topics-and-geography/topics/airline-time-performance-and-causes-flight-delays

Wikipedia. (2020, October 05). Alaska Air Group. Retrieved October 11, 2020, from https://en.wikipedia.org/wiki/Alaska\_Air\_Group