**Written Analysis and Graphs**

**Milestone 3**

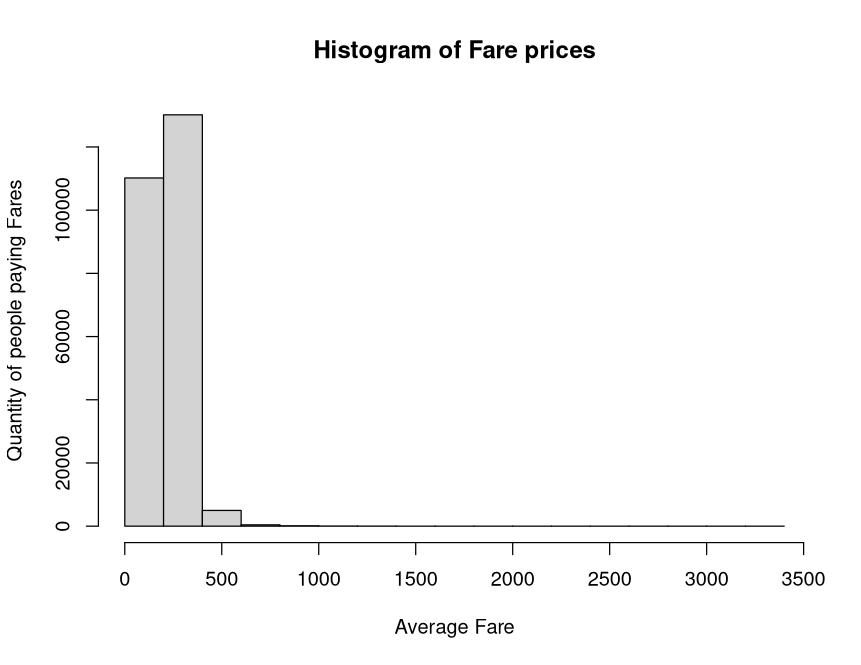
**Plot 1**

**A graph of a graph of distance between airports

AI-generated content may be incorrect.**

For this plot we can see that it is a right skewed plot indicating that most of the data would be on the left side of the histogram. Since a domestic flight only stays in their country of origin it wouldn’t make sense for the data to exceed 2892 miles (the distance of the USA from the west coast to the east coast). We can see that as the distance increases the number of flights decreases, which shows that people generally take more local flights rather than cross-country flights.

**Plot 2**

****

This plot demonstrates the average fare prices for every airline from the years 1993 to 2024. This is a right skewed plot and doesn’t show too much information but it’s for a reason. According to the Bureau of Transportation Statistics (BTS), the average fare price for a domestic flight in the US is about $385 per person. With this information we can say that the plot is accurate because most of the data within it falls between the range of $0 to $500. An outlier on the plot starts getting to the $1000 mark which can be explained by numerous factors, one being that the flight could be a cross-country flight (longer distance would mean more expensive fare) or the passenger took a more expensive airline.

**Plot 3**

**A graph of passengers

AI-generated content may be incorrect.**

This plot shows the number of passengers on different domestic airlines. This is a right skewed plot and most of the data points towards the numbers 0-500 passengers. The data has major outliers going all the way to 8000 passengers on a flight, which isn’t realistic and could be an error. This can be explained by the data only taking in the sampled passengers rather than the exact amount in each quarter of the year. For flight paths that aren’t as popular they tend to have less passengers than a flight with high volume to major popular cities, for example, a flight from LAX (Los Angeles, CA) to JFK (Queens, NY) would have a higher passenger count than a flight from ABQ (Albuquerque, NM) to TPA (Tampa, FL)

**Plot 4**

**A graph of a number of years

AI-generated content may be incorrect.**

This plot shows the number of flights taken yearly from the years 1993-2024. This is a uniform distribution because most of the data points are relatively close to each other with a few years having less data. We can see from the plot that during the years 1993-1995 flights taken were significantly lower than in the future years. Flights from the years 1997-2018 have stayed at a constant number with very few changes in value. If we look at the years 2020-2024, we can see a drop in flights, this can be explained by the COVID-19 pandemic when air travel was at an all-time low worldwide. Flights from the years 1997-2018 have had a constant number of

**Milestone 4**

**Plot 5**

A graph of passengers

AI-generated content may be incorrect.

The data that had to be removed was the enormous outliers on the histogram for the number of passengers on a plane. The previous plot had values extending to over 8000 passengers which isn’t realistic on a singular flight. The new histogram above displays a number that is still large but has gotten to the maximum which is 1100. The highest number of people ever on a singular flight in modern history is 1088 passengers from Addis Ababa to Israel. With this change to the histogram, it gives a more accurate look into how many people realistically were on US flight routes throughout the years and shows that most values are in the 0 – 100 range.

**Milestone 5**

**Calculated in R**

Mean, Median, Variance, and Standard Deviation for Plot 2 (Histogram of Fare Prices):

Mean = 218.9796

Median = 209.32

Variance = 6785.226

Standard Deviation = 82.37249

Mean, Median, Variance, and Standard Deviation for Plot 5 (Histogram of Trimmed Data for Passengers):

Trimmed Mean = 190.7183

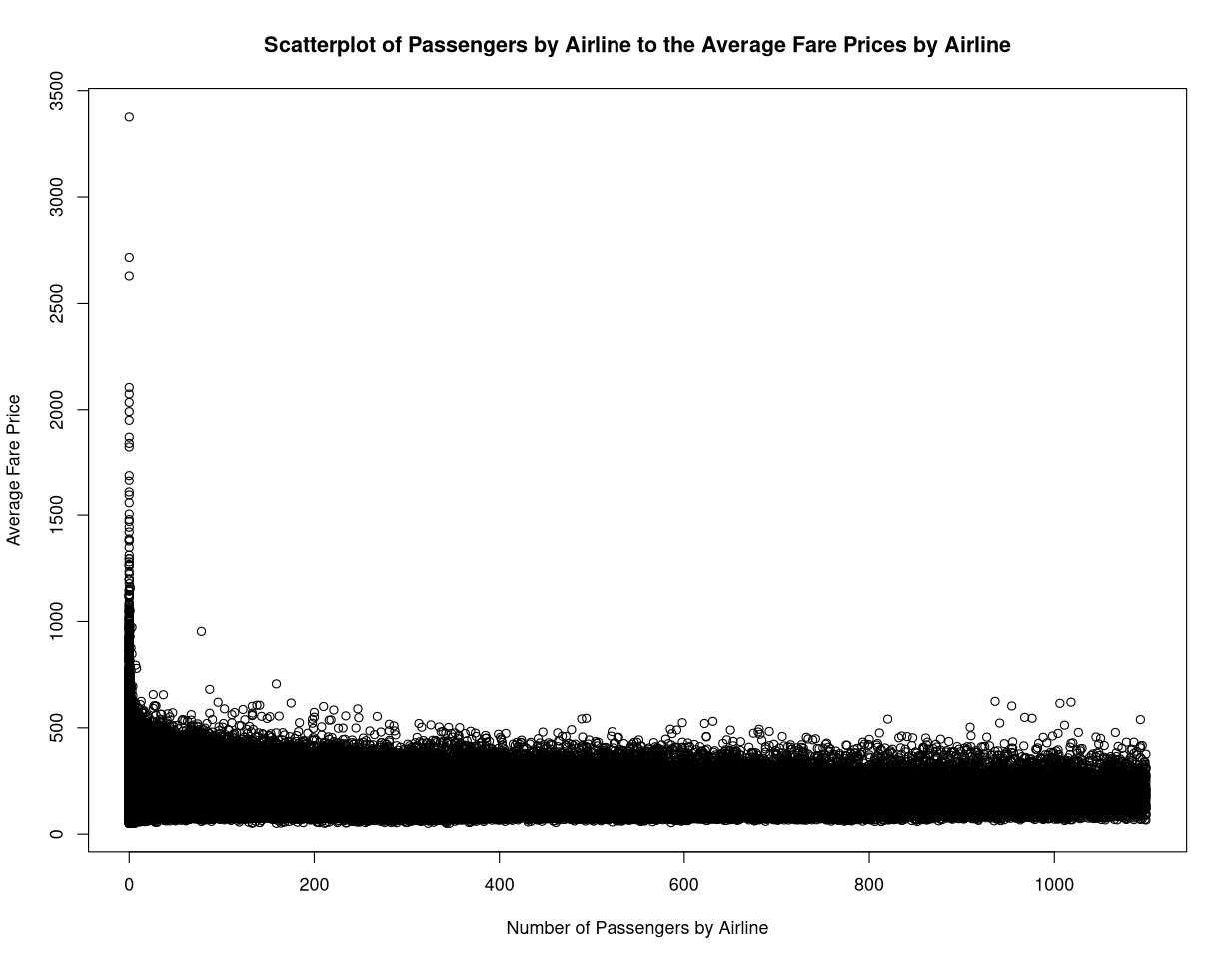
Median = 95

Variance = 56525.9

Standard Deviation = 237.7518

The mean and median for Plot 5 have two very different values and the reason for that is the data is skewed extremely right. For this data column I took the trimmed mean instead of the regular mean because the maximum number of passengers was an outrageous outlier of 8200 passengers which isn’t realistic. Trimming this plot made it so that the data will be more precise and give an accurate answer when asking for certain statistics (mean, median, variance, and standard deviation). Plot 2 has a very similar appearance to Plot 5 since it has an extremely right skew as well, even more than Plot 5. The difference with the two plots is that Plot 5 has data points from start to finish going down exponentially while Plot 2 has data points mainly from the 0-750 range in a maximum of 3500, these can be shown through the standard deviation. The standard deviations are the values that help us understand the spread of the data from its mean and for Plot 2 we can see that the spread isn’t that large therefore showing a lower deviation of **82.37249**. For Plot 5, however, we can see a lot of data from start to end and it is reflected by the larger standard deviation of **237.7518**.

**Milestone 6**

****

**The correlation between the number of passengers and the average fare price is -0.2086873**

Due to the passenger variable being a trimmed version of the data, few of the fare price variables have big outliers but judging from the rest of the scatterplot most values fall under 500 and have a steady negative rate of change. As the number of passengers increases the average fare decreases slightly. The two variables are closely related because they stay at around 500 for most of the data