|  |  |  |
| --- | --- | --- |
| Biggening, I wanted to see the columns available  Fl\_Date,OP\_Carrien\_name, OP\_Carriel\_FL\_NM ,  Weather delay  Nas Delay  Security Delay  Late\_aircraft\_Delay Just by looking at these results we an see there are latedelys flys from 51 to 108 minutes  I defined by variable y as arrival delay |  | These are my variables X |
| I wanted to find the top 5 airlines | top\_airlines = df['OP\_CARRIER\_NAME'].value\_counts().head(5)# AI i did ask ai what was the fastest function to find top5  print(top\_airlines) |  |
| Next I did a correlation matrix between the variables and noticed dep delay | correlation\_matrix = df[['ARR\_DELAY', 'DEP\_DELAY', 'CARRIER\_DELAY', 'LATE\_AIRCRAFT\_DELAY', 'NAS\_DELAY', 'WEATHER\_DELAY']].corr().round(2)  print(correlation\_matrix)  sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')  plt.show() | Dep\_delay has .96 which that is a strong relationship  Carrier\_Delay .57 some relation might be luck but its still stroong  And late \_aircraft delay is .52 still some relationship |
| Now creating a OLS model using Y as my independent variable and my xls these was the statistical result  Using the print(model.params)  Used to print the regression coefficients |  | What this correlation is saying is that by every minute it departs late its delayed .93 min  Each min of airline related it delays.07 if the last airflight was late for a minute its delayed extra .06  Each minute weather delay it ads .10 |
| Now plotting the OLS regression | fig, ax = plt.subplots()  fig = sm.graphics.plot\_fit(model, 1, ax=ax)  plt.title("OLS Fit: DEP\_DELAY vs ARR\_DELAY")  plt.xlabel("Departure Delay")  plt.ylabel("Arrival Delay")  plt.show() | This graph shows that the later a plane takes off, the later it usually arrives. The red line shows what the model predicts, and it follows the blue dots (real data) really well. This means departure delay is a strong reason why flights arrive late. A few flights are way off the line — those are rare cases with big delays. |
|  |  |  |

OVERALL SUMMARY

In this analysis, we used flight data to understand what causes arrival delays. We focused on the top 5 airlines with the most flights and built a regression model to predict how late a flight arrives based on four types of delays: departure, carrier-related, weather, and late incoming aircraft.

The results showed that departure delay is the strongest predictor of arrival delay. In simple terms, if a flight takes off late, it will almost always arrive late too. Other factors like weather and airline issues also contribute, but not as much. The model was statistically strong, with all predictors being significant and a good fit between the predicted and actual delays. Visuals like scatter plots and a fitted line confirmed the pattern and supported the model's accuracy.

Overall, the analysis shows that to reduce arrival delays, airlines should focus on minimizing departure delays first. The data proves that getting flights off the ground on time is the best way to keep arrivals on time.