Project Group - 2

A stormy flight: Investigating the correlation between weather patterns and delays on the JFK-LAX route

Members: Nynke Leppink, Alexander du Marchie Sarvaas, Samuel du Marchie Sarvaas, Sanne van Proosdij, Emma Toet

Student numbers: 5070988, 5547156, 4802829, 5326990, 4849892

Introduction

Flight delays are a common source of irritation for many travellers. Factors such as weather and technical issues frequently cause these disruptions. For frequent flyers, dealing with delays has become an unfortunate norm. This report will explore flight delays on the route between John F. Kennedy International Airport (JFK) and Los Angeles International Airport (LAX), which is one of the busiest air routes in the world, with over 150 flights per week. LAX ranks among the top ten busiest airports globally [1]. The report will examine how weather conditions at JFK and LAX impact flight delays on this route.

[1] https://www.airport-technology.com/features/the-top-10-busiest-airports-in-the-world/

Research objective

For this research the following question will be answered: "How do weather conditions at JFK (New York) and LAX (Los Angeles) affect flight delays on this flight route?"

To answer this question a dataset of flight information from the year 2013 will be used [2]. This dataset contains flight information of different flight routes in the United States. This research focuses on the flight route between John F. Kennedy International Airport (JFK) in New York and Los Angeles International Airport (LAX). The information in this dataset can be used to analyse the departure and arrival delays of these airports.

For the weather conditions at JFK and LAX the weather information of Iowa Environmental Mesonet will be used [3]. On this website different weather variables are available and will be combined in one dataset.

Sub questions

1. What is the relationship between (severe) weather patterns and the duration and occurence of flight delays, and how does this vary depending on the point of departure and arrival? \

2. Which specific weather conditions (e.g., visibility, wind speed and wind gust, sky level coverage, precipitation and pressure altimeter) most significantly impact flight delays on the JFK-LAX route? \
3. How does the time of year affect the relationship between weather conditions and flight delays (and how can this be used for more accurate future delay predictions)? \
4. How do the operational performances of various airlines, such as their on-time records, differ in response to weather disruptions along the JFK-LAX route? \
5. Are there unexpected patterns in flight delays based on the time of day and specific weather conditions, such as for example increased delays during morning rush hours in bad weather? Or were there unexpected delays due to factors not having anything to do with weather. \

The information of both datasets will be analysed to answer the research- and the sub question.

References: [2] https://www.kaggle.com/datasets/mahoora00135/flights?resource=download
[3] https://mesonet.agron.iastate.edu/request/download.phtml?network=NY_ASOS#

Conditions and restrains

In this study, there are several constraints that influence the analysis and findings. One of the most important factors is weather conditions, which include a variety of elements that are important to understanding flight operations. The study will consider factors such as visibility, wind speed and wind gust, sky level coverage, precipitation and pressure altimeter. The to-be-used datasat has five airlines operating this route: United Airlines Inc., Virgin America, JetBlue Airways, American Airlines Inc., and Delta Air Lines Inc. By investigating the performance of these airlines, the aim is to identify trends and correlations between weather conditions and flight operations. This will enable a better understanding of how various factors impact delays throughout the year.

The temporal scope of the study covers the year 2013, which provides a big dataset for analysis. The dataset (consisting of more than 11.000 flights) has flights at an hourly level, monitoring approximately 30 flights each day.

Geographically, the analysis will concentrate on the route between two major international airports: John F. Kennedy International Airport (JFK) in New York and Los Angeles International Airport (LAX). By focusing on these key locations, the weather conditions specific to each airport will be explored.

Quantitative Analysis

To better understand the impact of weather on flight delays between JFK and LAX three statistical analyses will be performed.

Descriptive Statistics: A first objective is to calculate the mean of flight delays on clear days compared to bad weather days. This will provide insights into how different weather conditions can affect delay durations.
Correlation Analysis: A second objective is assessing the relationship between various weather variables, such as wind speed and precipitation, and the duration of flight delays. This analysis can help identify which weather factors have the strongest influence on delays. (In case we want to expand with a predictive model).
Regression Analysis: To predict flight delays, multiple linear regression or logistic regression analyses will be conducted. These models will consider multiple factors, including weather conditions and the day of the week, to determine their combined effect on delay occurrence and duration.

Visualisation

Here, different ways of translating the findings into comprehensible visualisations will be used. Maps: Heatmaps showing concentration of delays by geographic area. Graphs:

1. Line graphs to display delays over time, categorised by weather types (e.g., delays during rain vs. delays during clear days). \
2. Bar charts showing the average delay duration by month or airline. \
3. Scatter Plots: To visualise correlations between variables (e.g., wind speed vs. delay duration).

Contribution Statement

Be specific. Some of the tasks can be coding (expect everyone to do this), background research, conceptualisation, visualisation, data analysis, data modelling

For the proposal we collabarated together.

Author 1:

Author 2:

Author 3:

Author 4:

Author 5:

Data Used

The datasets used are described in the research objective

Data Pipeline

The dataframe for the weather variables at JFK are modified in tbe following cell and the output of the dataframe is given.

In [1]:
```
import pandas as pd
import numpy as np
weather_JFK = pd.read_csv('JFK.csv')
weather_JFK[['Date', 'Time']] = weather_JFK['valid'].str.split(' ', expand=True)


weather_JFK = weather_JFK.drop(['tmpf', 'dwpf', 'relh', 'drct', 'mslp', 'skyl1',
                                "skyl2",  'skyl3', 'skyl4',  'wxcodes', 'feel',
                                'ice_accretion_6hr', 'peak_wind_gust', 'peak_win

weather_JFK = weather_JFK[['station', 'Date', 'Time', 'sknt', 'gust', 'p01i', 'a
weather_JFK.rename(columns={
    'sknt': 'Wind Speed [in knots]',
    'gust': 'Wind Gust [in knots]',
    'vsby': 'Visibility [in miles]',
    'p01i': 'One hour precipitation [in inches]',
    'alti': 'Pressure altimeter [in inches]',
    'skyc1': 'Sky Level 1 Coverage',
    'skyc2': 'Sky Level 2 Coverage',
    'skyc3': 'Sky Level 3 Coverage',
    'skyc4': 'Sky Level 4 Coverage'
}, inplace = True)

weather_JFK = weather_JFK[weather_JFK['Time'].str.endswith(':51')]
display(weather_JFK)
```

	station	Date	Time	Wind Speed [in knots]	Wind Gust [in knots]	One hour precipitation [in inches]	Pressure altimeter [in inches]	Visibility [in miles]	Sky Level 1 Coverage
0	JFK	2013-01-01	00:51	14.0	NaN	0.00	29.96	10.0	OVC
1	JFK	2013-01-01	01:51	15.0	NaN	0.00	29.93	10.0	OVC
2	JFK	2013-01-01	02:51	17.0	NaN	0.00	29.92	10.0	OVC
3	JFK	2013-01-01	03:51	16.0	NaN	0.00	29.92	10.0	OVC
4	JFK	2013-01-01	04:51	13.0	NaN	0.00	29.92	10.0	OVC
...	...	...	...	...	...	...	...	...	...
9828	JFK	2013-12-31	19:51	12.0	NaN	0.00	30.09	10.0	FEW
9829	JFK	2013-12-31	20:51	16.0	NaN	T	30.12	9.0	BKN
9830	JFK	2013-12-31	21:51	18.0	26.0	T	30.13	10.0	FEW
9831	JFK	2013-12-31	22:51	18.0	24.0	0.00	30.15	10.0	FEW
9832	JFK	2013-12-31	23:51	15.0	24.0	0.00	30.19	10.0	FEW

8747 rows × 12 columns

The dataframe for the weather variables at LAX are modified in tbe following cell and the output of the dataframe is given.

In [2]:

```
weather_LAX = pd.read_csv('LAX.csv')
weather_LAX[['Date', 'Time']] = weather_LAX['valid'].str.split(' ', expand=True)

weather_LAX = weather_LAX.drop(['tmpf', 'dwpf', 'relh', 'drct', 'mslp', 'skyl1',
                                "skyl2",  'skyl3', 'skyl4',  'wxcodes', 'feel',
                                'ice_accretion_6hr', 'peak_wind_gust', 'peak_win

weather_LAX = weather_LAX[['station', 'Date', 'Time', 'sknt', 'gust', 'p01i', 'a
weather_LAX.rename(columns={
    'sknt': 'Wind Speed [in knots]',
    'gust': 'Wind Gust [in knots]',
    'vsby': 'Visibility [in miles]',
    'p01i': 'One hour precipitation [in inches]',
    'alti': 'Pressure altimeter [in inches]',
    'skyc1': 'Sky Level 1 Coverage',
    'skyc2': 'Sky Level 2 Coverage',
    'skyc3': 'Sky Level 3 Coverage',
    'skyc4': 'Sky Level 4 Coverage'
}, inplace = True)


weather_LAX = weather_LAX[weather_LAX['Time'].str.endswith(':53')]
display(weather_LAX)
```

	station	Date	Time	Wind Speed [in knots]	Wind Gust [in knots]	One hour precipitation [in inches]	Pressure altimeter [in inches]	Visibility [in miles]	Sky Level 1 Coverage
0	LAX	2013-01-01	00:53	8.0	NaN	0.00	30.21	10.0	FEW
1	LAX	2013-01-01	01:53	7.0	NaN	0.00	30.21	10.0	FEW
2	LAX	2013-01-01	02:53	4.0	NaN	0.00	30.21	10.0	CLR
3	LAX	2013-01-01	03:53	5.0	NaN	0.00	30.23	10.0	FEW
4	LAX	2013-01-01	04:53	3.0	NaN	0.00	30.24	10.0	FEW
...	...	...	...	...	...	...	...	...	...
10042	LAX	2013-12-31	19:53	7.0	NaN	0.00	30.15	10.0	FEW
10043	LAX	2013-12-31	20:53	6.0	NaN	0.00	30.12	10.0	FEW
10045	LAX	2013-12-31	21:53	6.0	NaN	0.00	30.11	10.0	FEW
10046	LAX	2013-12-31	22:53	7.0	NaN	0.00	30.10	10.0	FEW
10047	LAX	2013-12-31	23:53	7.0	NaN	0.00	30.10	10.0	FEW

8741 rows × 12 columns

The flight dataset has been modified from its original form, filtering it to only include the flights from JFK to LAX. Some unnecesary columns have been removed, and the time have been cleaned. The original dataset cannot be uploaded in github because of the size, only the modified version is uploaded.

In [3]:
```
import pandas as pd
flights = pd.read_csv('Vluchten_2013_dataset.csv')
NYLA = flights[(flights['origin'] == 'JFK') & (flights['dest'] == 'LAX')]
```

In [4]:
```
NYLA['date'] = pd.to_datetime(NYLA[['year', 'month', 'day']])
NYLA = NYLA.drop(['id', 'year', 'month', 'day', 'time_hour', 'distance', 'hour',
cols = ['date'] + [col for col in NYLA if col != 'date']
```

```
NYLA = NYLA[cols]
NYLA = NYLA.sort_values(by='date')
NYLA = NYLA.dropna(axis=0)
NYLA = NYLA.reset_index()
NYLA = NYLA.drop(['index'], axis=1)
```

```
C:\Users\nynke\AppData\Local\Temp\ipykernel_4568\1040830742.py:1: SettingWithCopy
Warning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  NYLA['date'] = pd.to_datetime(NYLA[['year', 'month', 'day']])
```

In [5]:

```
#function to convert departure and arrival times in floats, to datetime objects

def convert_to_time(time_value):
    time_value = int(time_value)

    if time_value < 100:  # If it's less than 100, it's only minutes
        return f"00:{time_value:02d}"
    else:
        time_str = str(time_value).zfill(4)
        hour = int(time_str[:2])
        minute = int(time_str[2:])
        return f"{hour:02d}:{minute:02d}"


timecolumns = ['dep_time', 'sched_dep_time', 'arr_time', 'sched_arr_time']

for col in timecolumns:
    NYLA[col] = NYLA[col].apply(convert_to_time)
```

In [6]:

```
NYLA.tail()
```

Out[6]:

	date	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay
11154	2013-12-31	07:24	07:00	24.0	11:03	10:25	38.(
11155	2013-12-31	06:59	07:00	-1.0	10:15	10:15	0.(
11156	2013-12-31	06:30	06:31	-1.0	09:55	09:58	-3.(
11157	2013-12-31	15:31	15:30	1.0	19:18	19:03	15.(
11158	2013-12-31	11:32	11:29	3.0	14:48	14:46	2.(

In [7]:

```
NYLA.to_csv('FLIGHT_csv_clean.csv', index=False)
```

In [ ]: