UK National Rail Analysis

•••

Presented by: Alex Kennedy, Arthur Levitan, Bryce Martin, Julia Snabes

Table of Contents

- Objective
- Background Info
- Building the Pipeline
- What is EDA?
- Business Questions
- Data Exploration & Insights
- Conclusion
- Appendix

Objective

Create a data pipeline to collect data from UK National Rail. Explore the UK National Rail dataset, generate business questions and gain insights into the train system based on data.

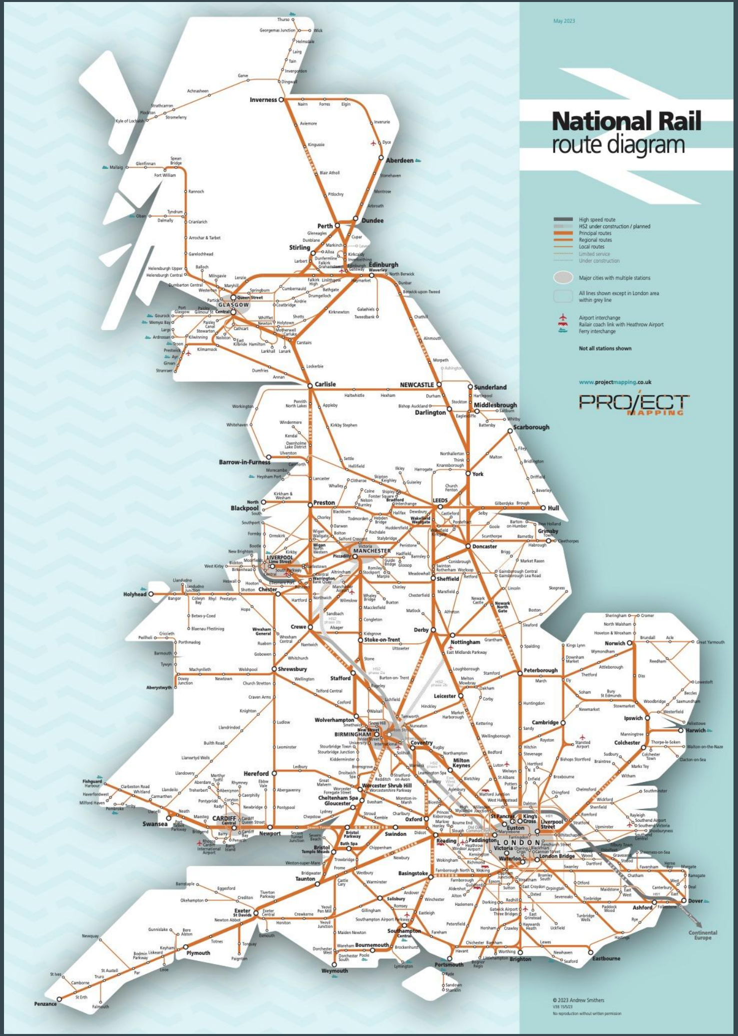
Background

-UK National Rail system

-Publically available data

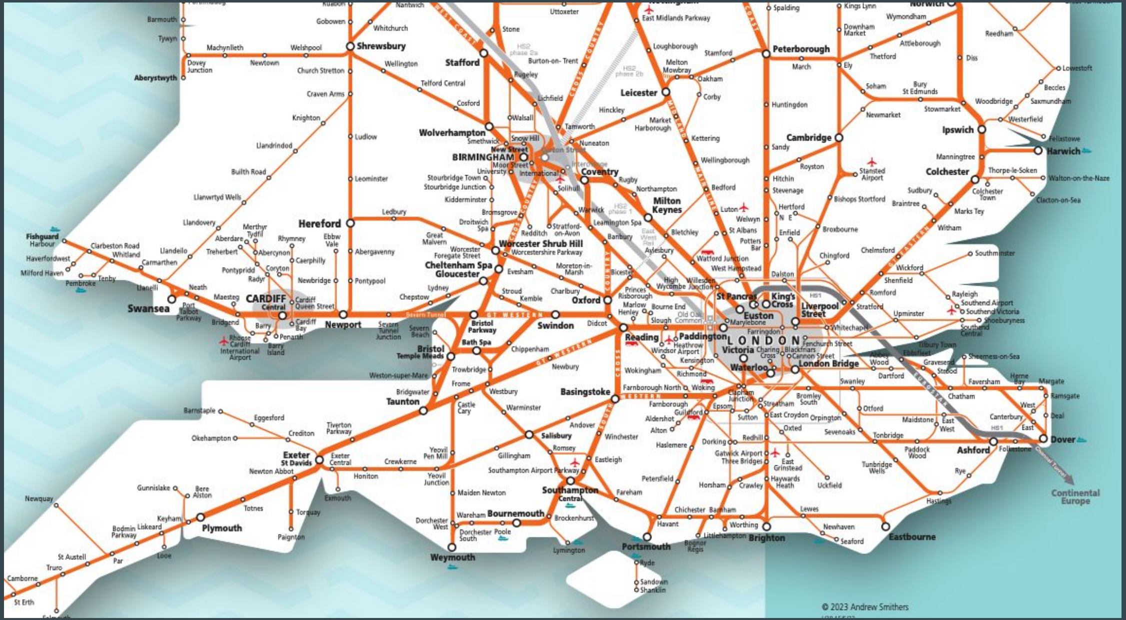
-Creation of, and changes to, train schedule records (scheduled arrivals/departures vs actual)

-Robust, highly used train system



Source: https://www.nationalrail.co.uk/travel-information/maps-of-the-national-rail-network/

Background



- ▬ High speed route
- ▬ HS2 under construction / planned
- ▬ Principal routes
- ▬ Regional routes
- ▬ Local routes
- ┈ Limited service
- ┈ Under construction
- ⬭ Major cities with multiple stations
- ▢ All lines shown except in London area within grey line
- ✈ Airport interchange
- Railair coach link with Heathrow Airport
- Ferry interchange

Not all stations shown

www.projectmapping.co.uk

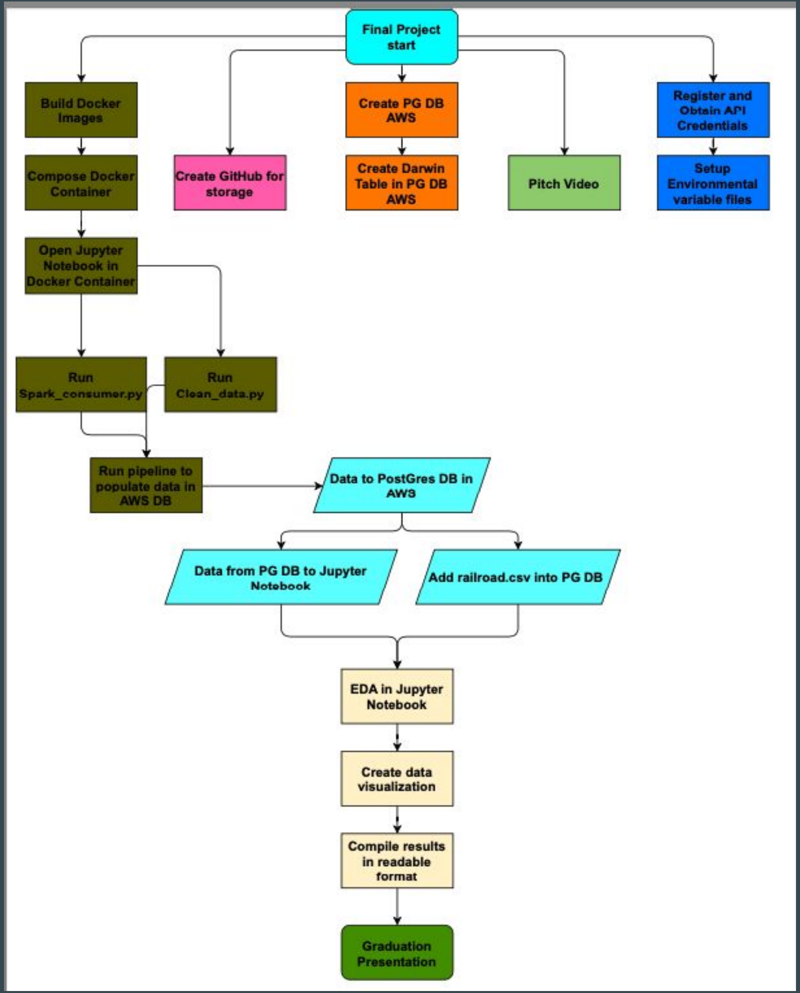
PROJECT MAPPING

Building the Data Pipeline

-Connected to a website that streams live train movement data

-Saved data to cloud due to large volume

-Completed exploratory data analysis on the data

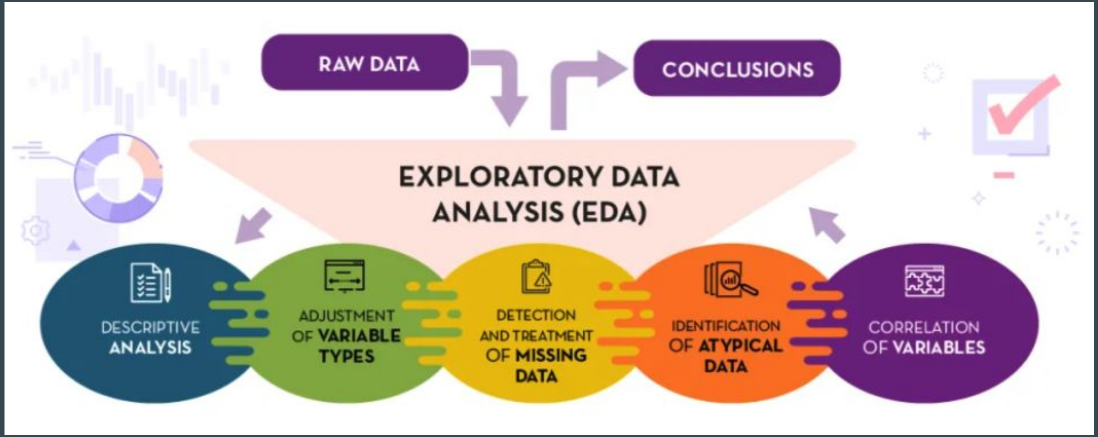

Final Project start
Build Docker Images
Compose Docker Container
Open Jupyter Notebook in Docker Container
Run Spark_consumer.py
Run Clean_data.py
Run pipeline to populate data in AWS DB
Create GitHub for storage
Create PG DB AWS
Create Darwin Table in PG DB AWS
Pitch Video
Register and Obtain API Credentials
Setup Environmental variable files
Data to PostGres DB in AWS
Data from PG DB to Jupyter Notebook
Add railroad.csv into PG DB
EDA in Jupyter Notebook
Create data visualization
Compile results in readable format
Graduation Presentation

What is EDA?

-Exploratory data analysis

-Analyzing and exploring a data set to draw meaningful insights

-Completed our EDA in Python, leveraging Pandas library


RAW DATA
CONCLUSIONS
EXPLORATORY DATA ANALYSIS (EDA)
DESCRIPTIVE ANALYSIS
ADJUSTMENT OF VARIABLE TYPES
DETECTION AND TREATMENT OF MISSING DATA
IDENTIFICATION OF ATYPICAL DATA
CORRELATION OF VARIABLES

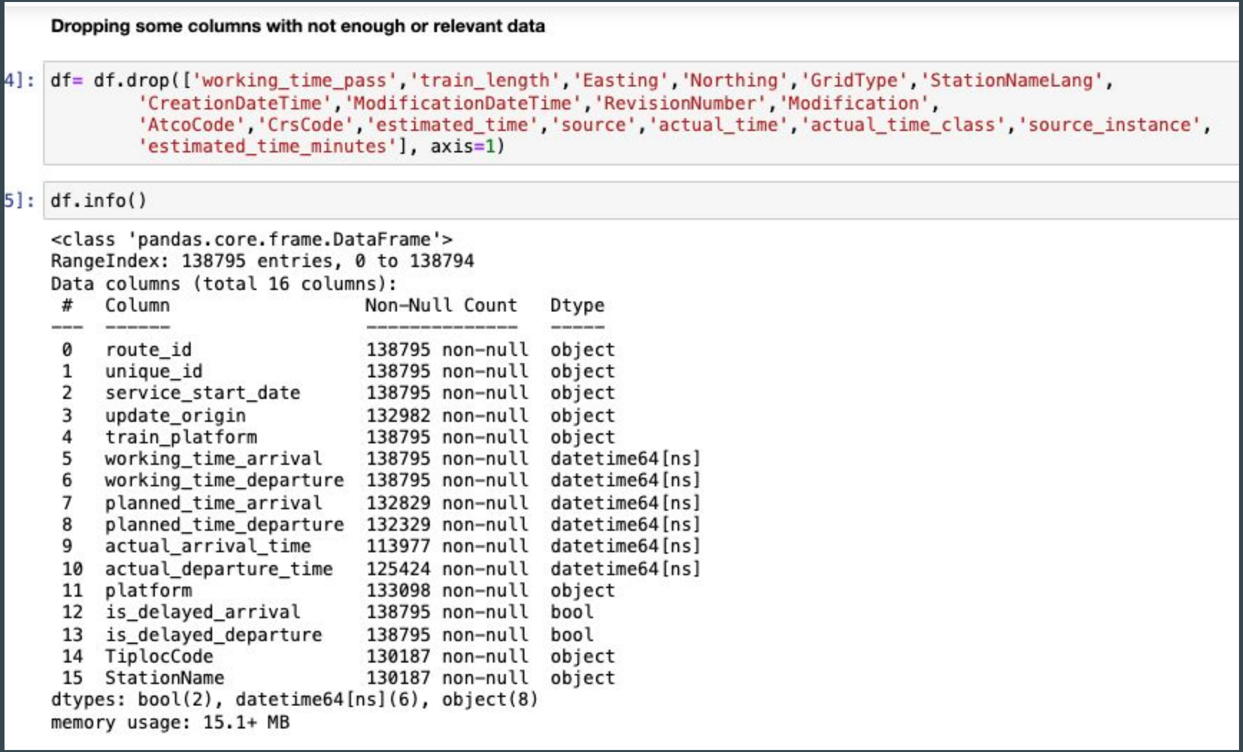
https://dev.to/ckawara/exploratory-data-analysis-ultimate-guide-3mea

EDA cont.

-Cleaning data

-Creating new columns

-Filtering data



Dropping some columns with not enough or relevant data

```
4]: df= df.drop(['working_time_pass','train_length','Easting','Northing','GridType','StationNameLang',
            'CreationDateTime','ModificationDateTime','RevisionNumber','Modification',
            'AtcoCode','CrsCode','estimated_time','source','actual_time','actual_time_class','source_instance',
            'estimated_time_minutes'], axis=1)
```

```
5]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 138795 entries, 0 to 138794
Data columns (total 16 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   route_id                138795 non-null  object
 1   unique_id               138795 non-null  object
 2   service_start_date      138795 non-null  object
 3   update_origin           132982 non-null  object
 4   train_platform          138795 non-null  object
 5   working_time_arrival    138795 non-null  datetime64[ns]
 6   working_time_departure  138795 non-null  datetime64[ns]
 7   planned_time_arrival    132829 non-null  datetime64[ns]
 8   planned_time_departure  132329 non-null  datetime64[ns]
 9   actual_arrival_time     113977 non-null  datetime64[ns]
 10  actual_departure_time   125424 non-null  datetime64[ns]
 11  platform                133098 non-null  object
 12  is_delayed_arrival      138795 non-null  bool
 13  is_delayed_departure    138795 non-null  bool
 14  TiplocCode              130187 non-null  object
 15  StationName             130187 non-null  object
dtypes: bool(2), datetime64[ns](6), object(8)
memory usage: 15.1+ MB
```

Business Questions

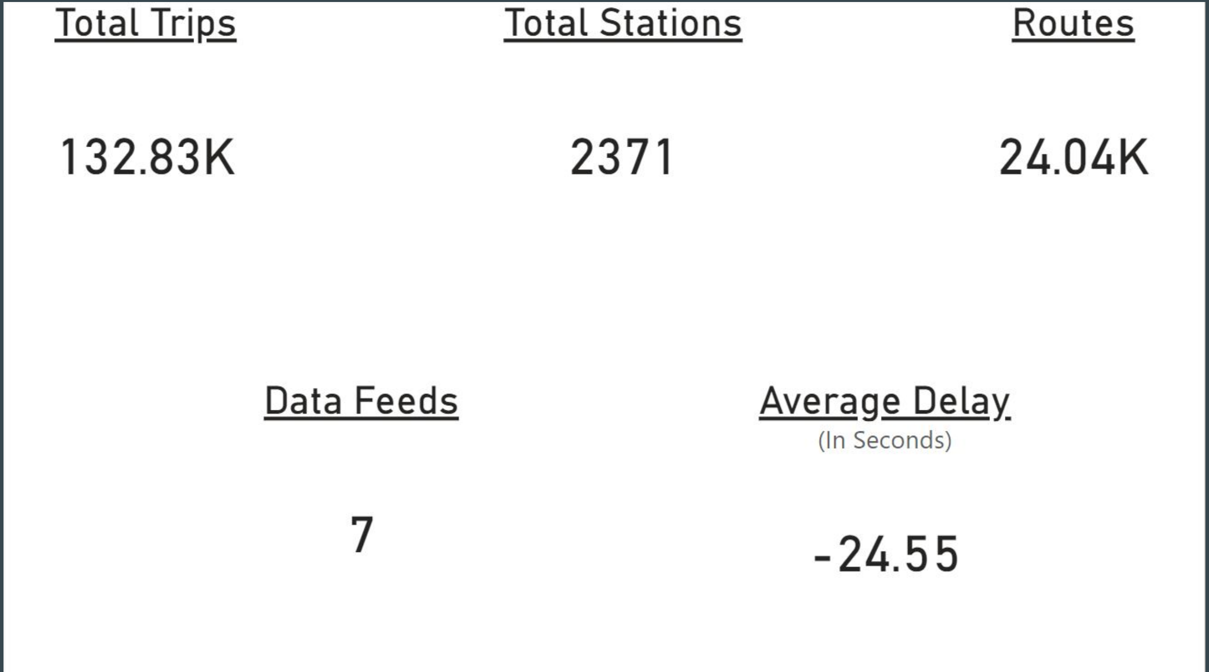
-How does the rail system appear to be operating based on data? How often are trains delayed?

-What are the relationships between working/planned times and actual times? Does National Rail need to adjust their scheduling?

-Are certain factors correlated to more frequent delays?

Dataset At-a-Glance

-Date range: 12/16/23 - 12/18/23



Total Trips

132.83K

Total Stations

2371

Routes

24.04K

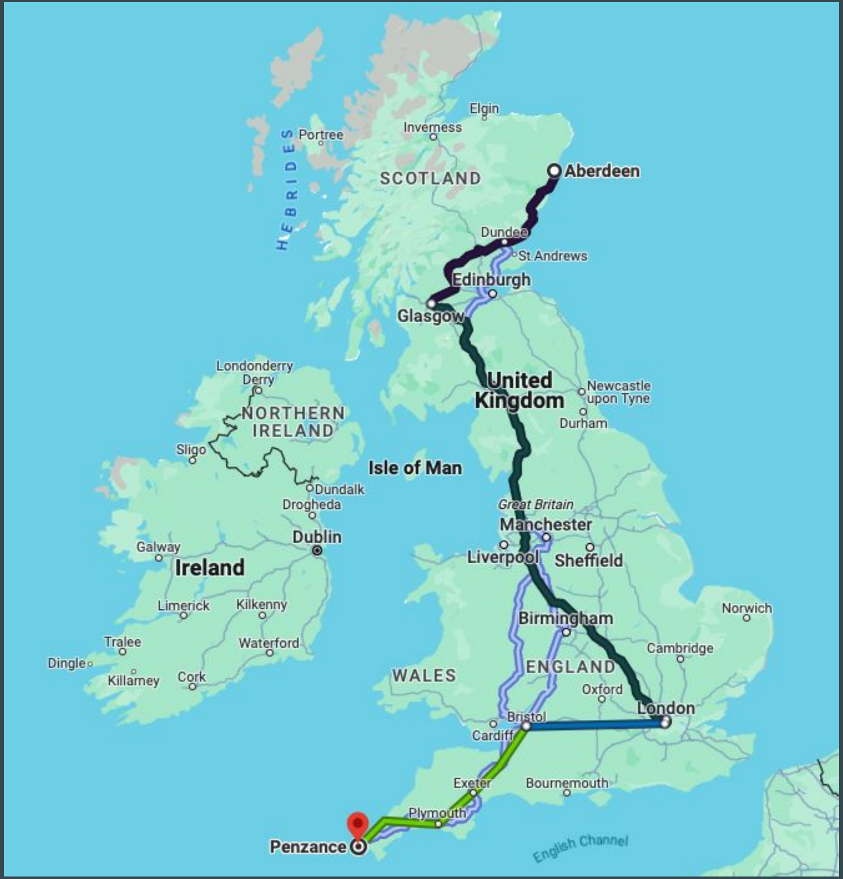
Data Feeds

7

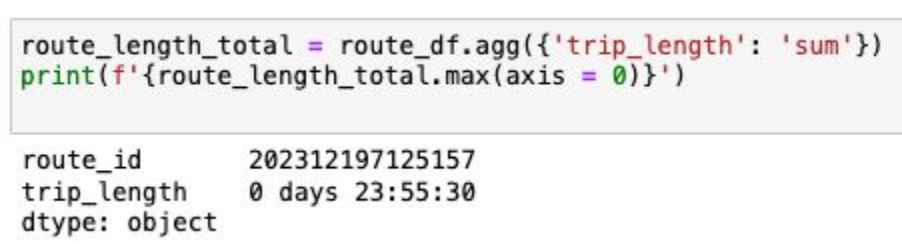
Average Delay
(In Seconds)

-24.55

What is the longest route based on time?



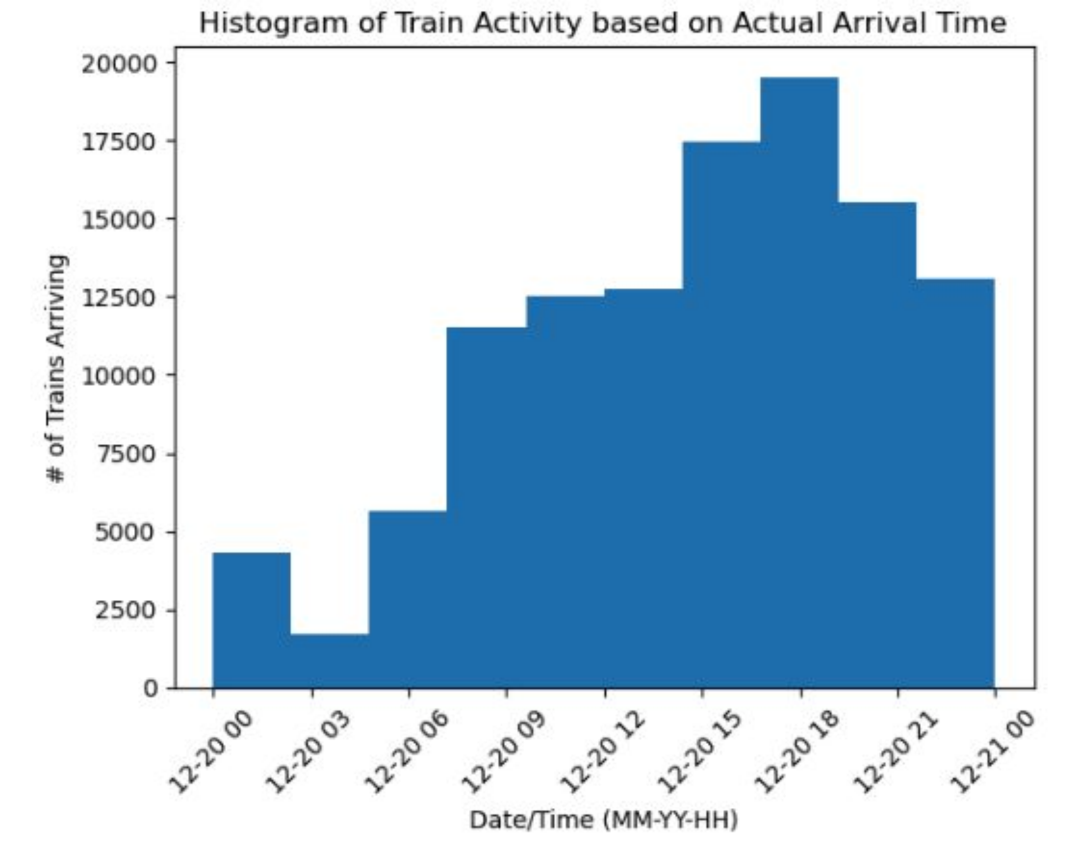
Aberdeen, Scotland →
Penzance, England



```
route_length_total = route_df.agg({'trip_length': 'sum'})
print(f'{route_length_total.max(axis = 0)}')
```

```
route_id       20231219712515 7
trip_length    0 days 23:55:30
dtype: object
```

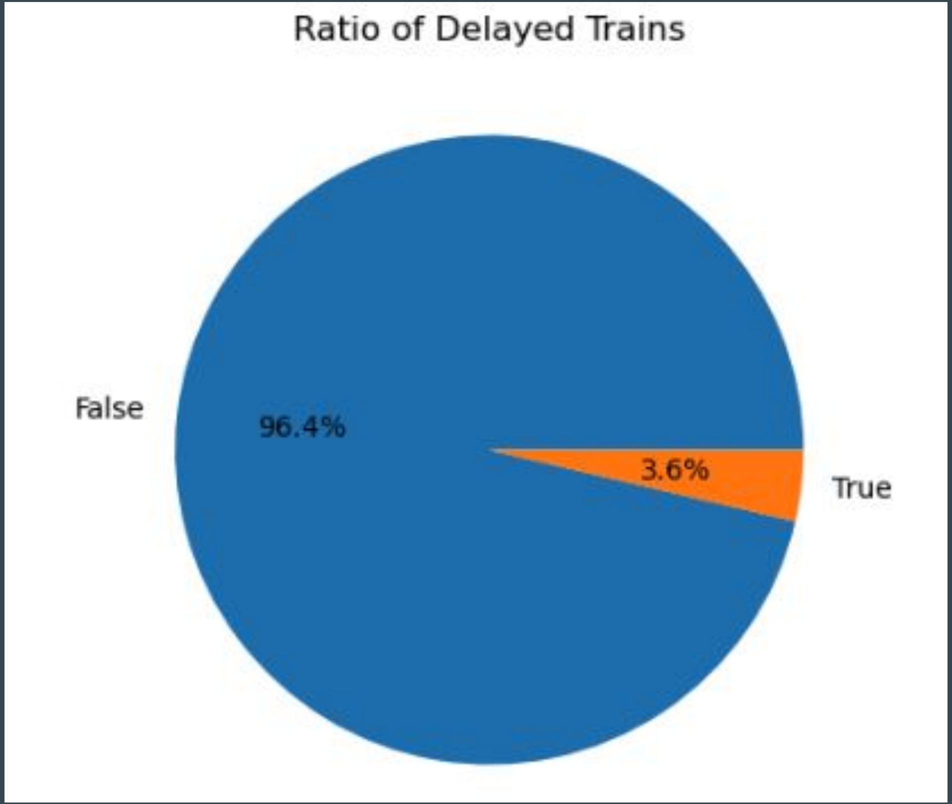
What times are stations busiest?



How often are trains delayed?

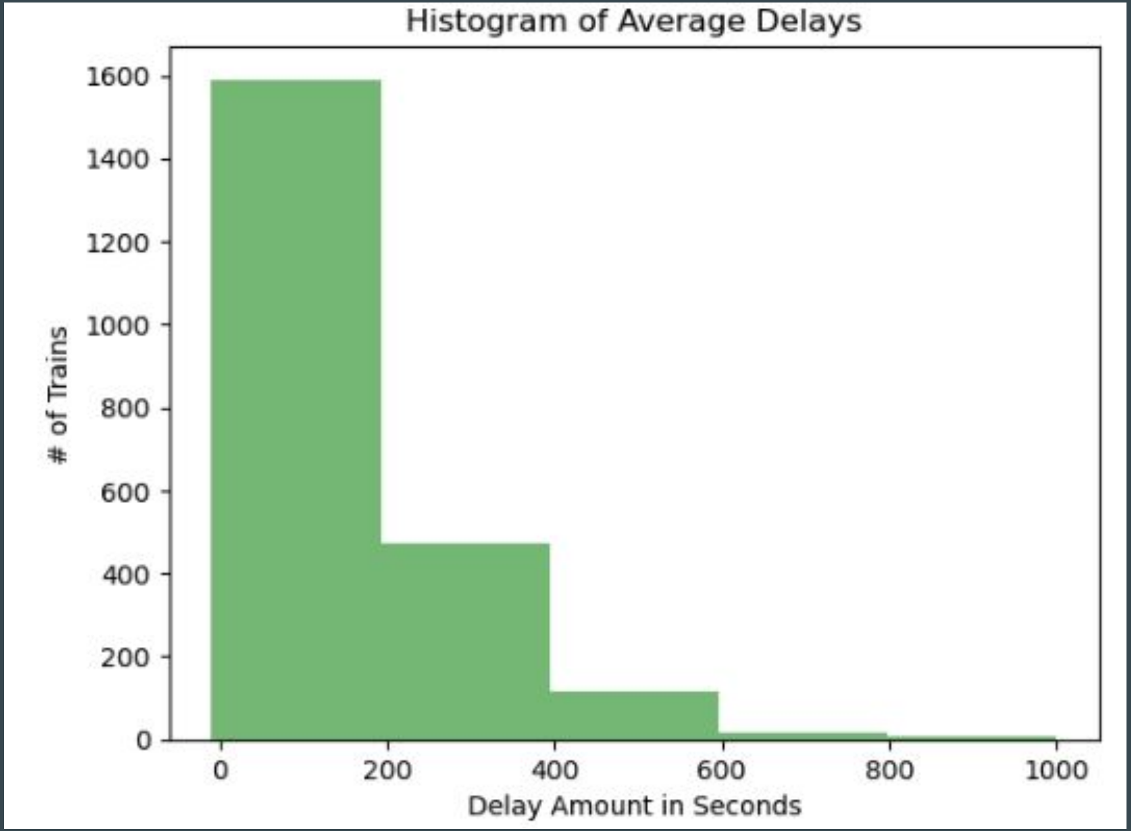
-**False**: (not delayed) 96.4% of the time

-**True**: (delayed) 3.6% of the time



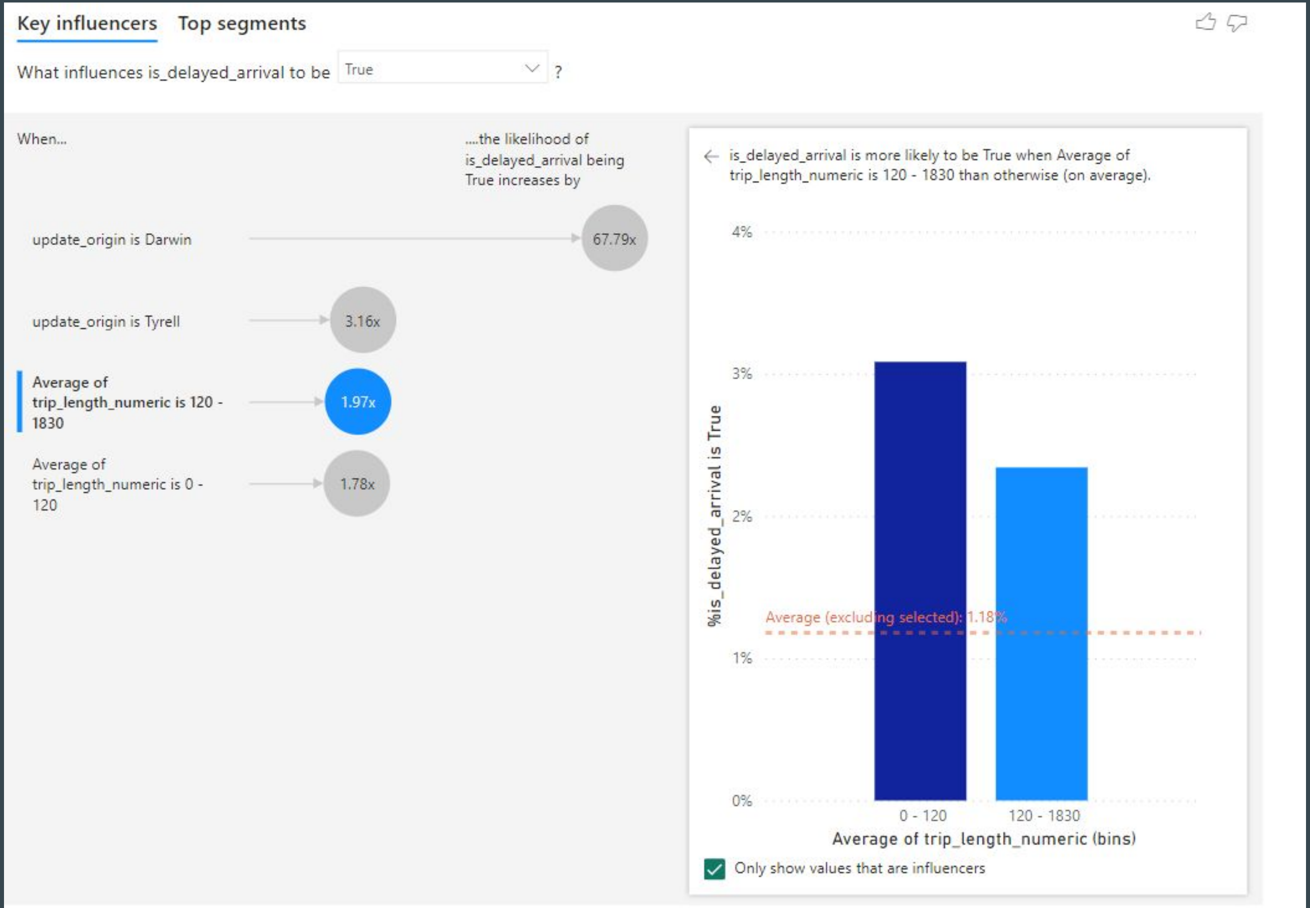
How long are most delays across all data?

-Majority of delays between 0-6 minutes



What key factors influence delayed arrival time?

-trip length found to be a key factor



What influences is_delayed_arrival to be True ?

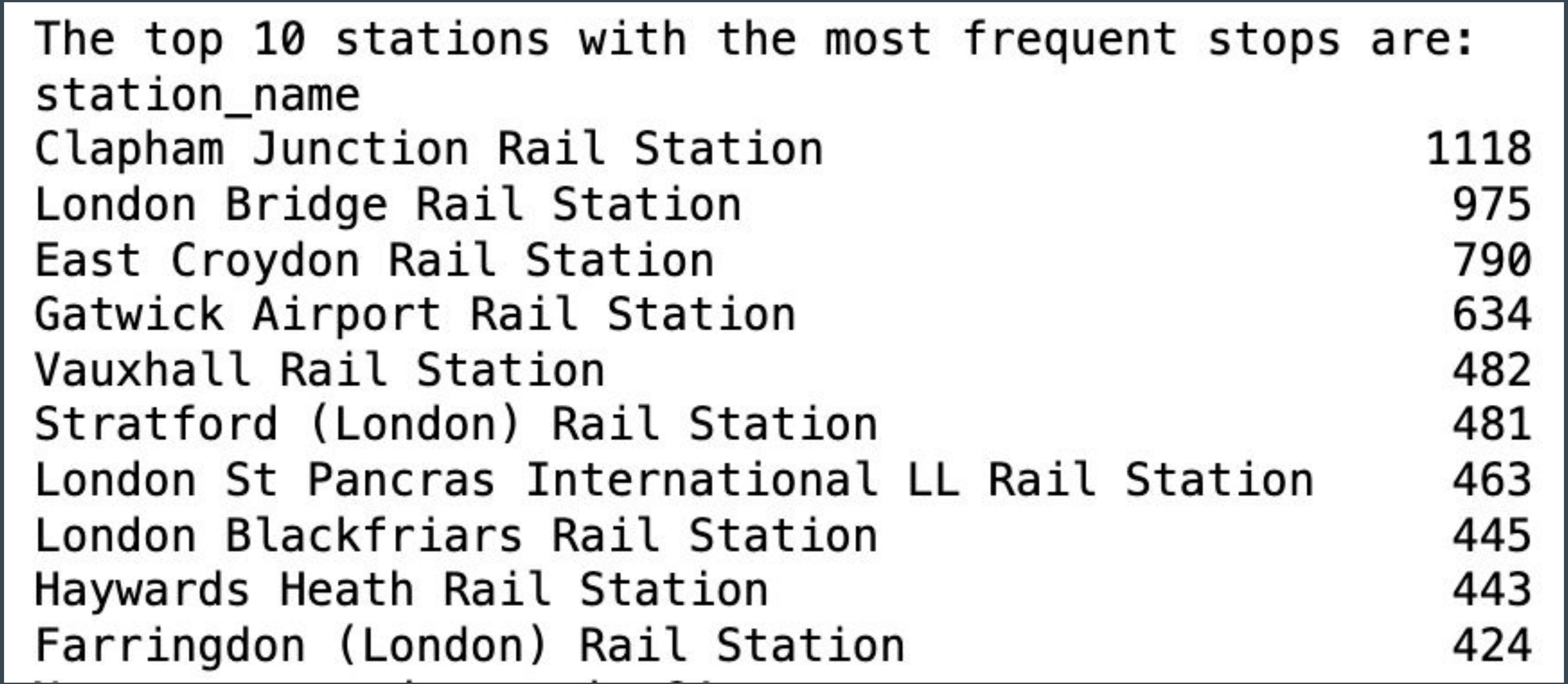
When...	....the likelihood of is_delayed_arrival being True increases by
update_origin is Darwin	67.79x
update_origin is Tyrell	3.16x
Average of trip_length_numeric is 120 - 1830	1.97x
Average of trip_length_numeric is 0 - 120	1.78x

is_delayed_arrival is more likely to be True when Average of trip_length_numeric is 120 - 1830 than otherwise (on average).

4%
3%
2%
1%
0%
%is_delayed_arrival is True
Average (excluding selected): 1.18%
0 - 120
120 - 1830
Average of trip_length_numeric (bins)

☑ Only show values that are influencers

What are busiest stations? Is delay amount affected by volume?



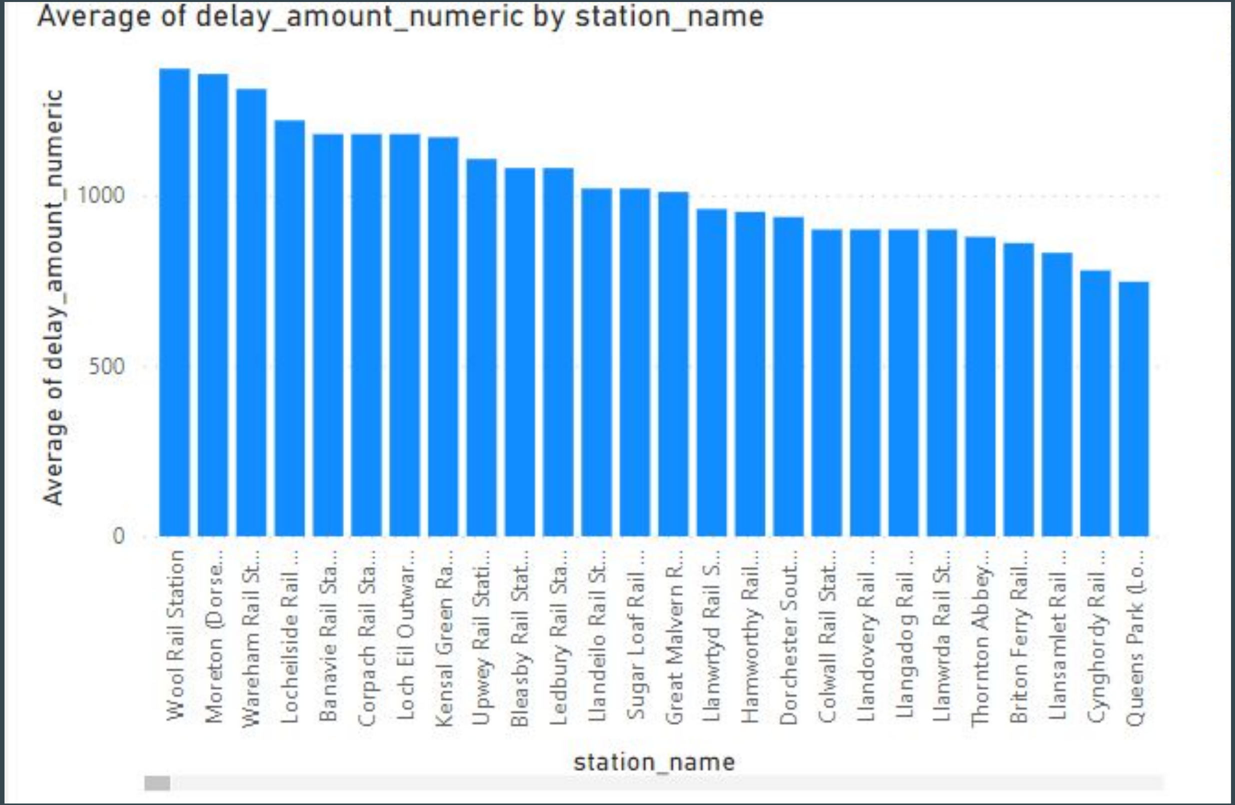
```
The top 10 stations with the most frequent stops are:
station_name
Clapham Junction Rail Station                           1118
London Bridge Rail Station                               975
East Croydon Rail Station                                790
Gatwick Airport Rail Station                             634
Vauxhall Rail Station                                    482
Stratford (London) Rail Station                          481
London St Pancras International LL Rail Station          463
London Blackfriars Rail Station                          445
Haywards Heath Rail Station                              443
Farringdon (London) Rail Station                         424
```

What are busiest stations? Is delay amount affected by volume?

-Average delay amount = in seconds

-Highest average delay amount:

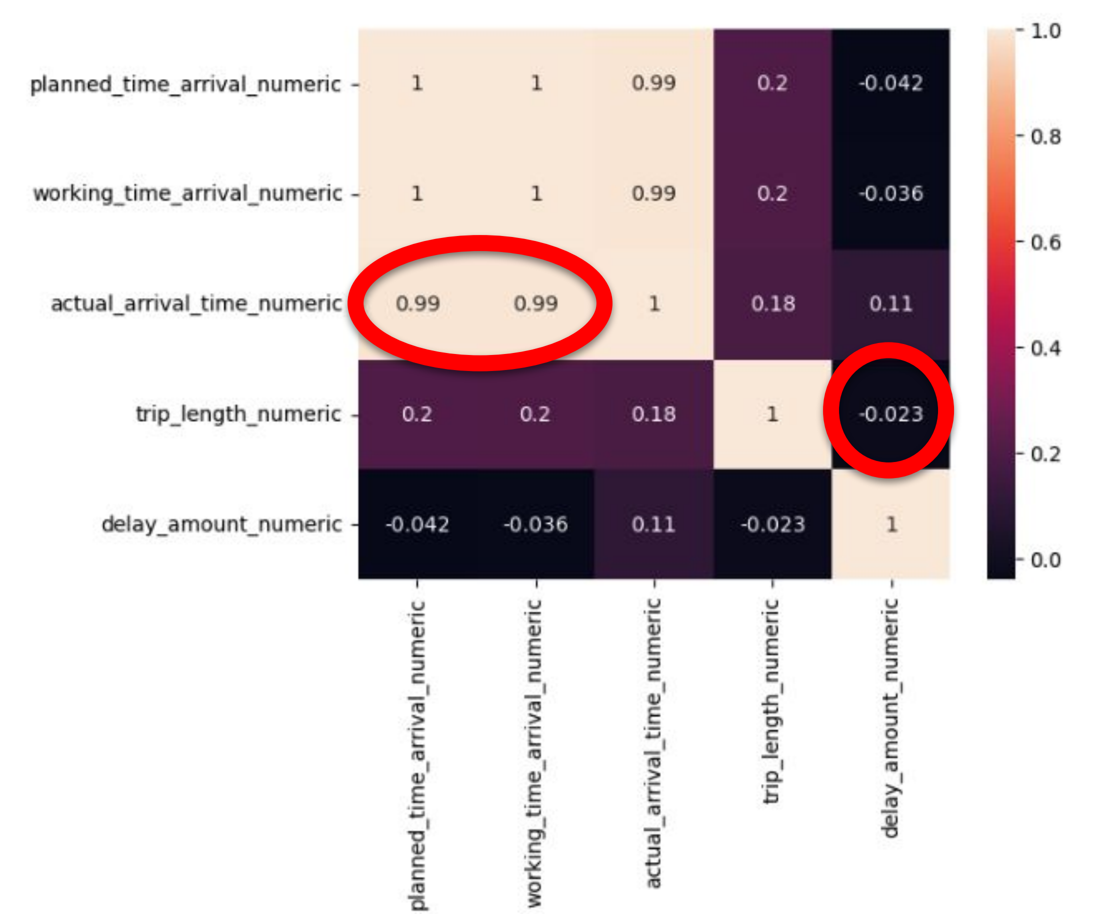
Wool Rail Station
1,371.43 seconds
(23 minutes)



Correlation matrix of time relationships

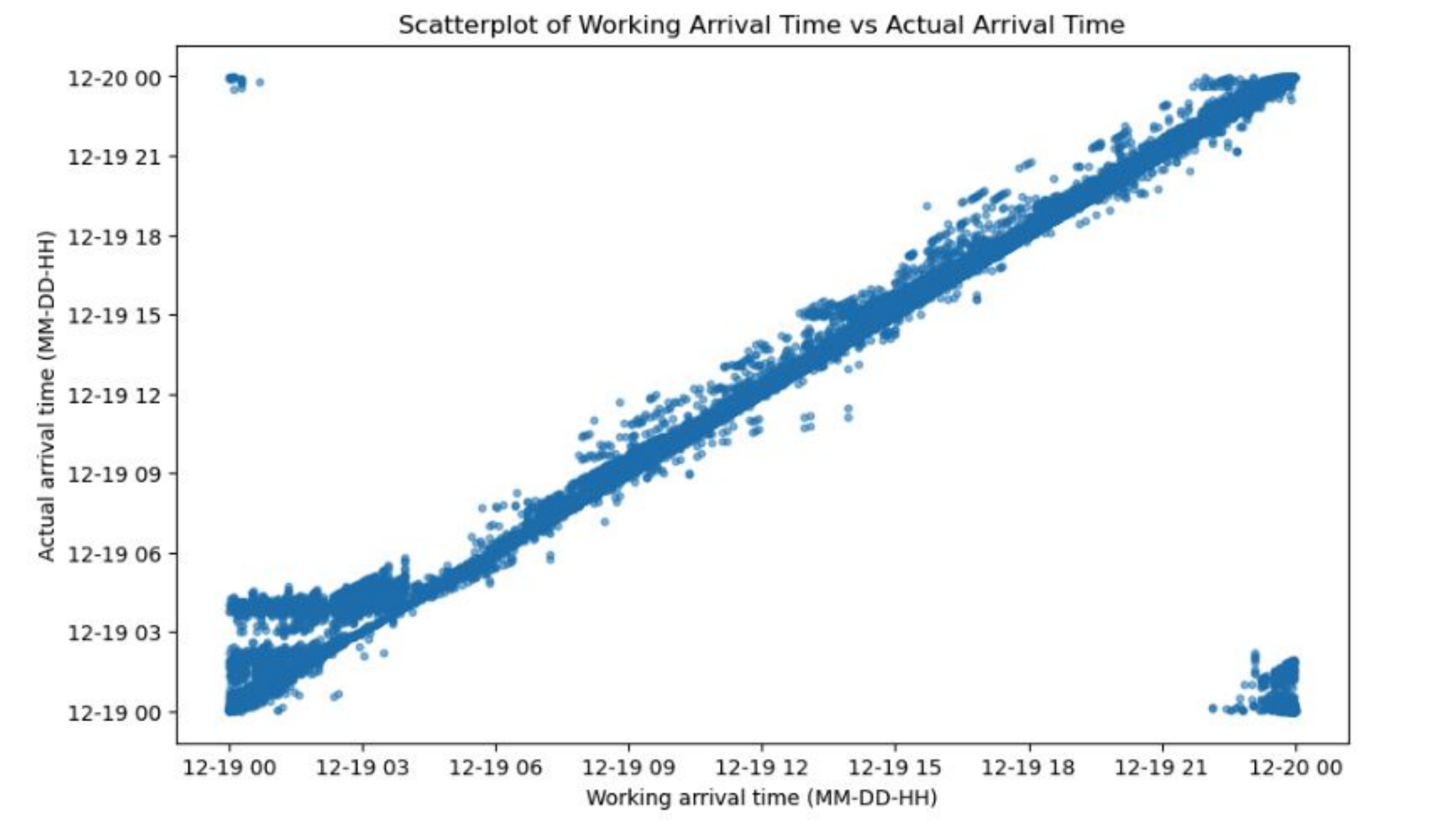
-*Strong negative correlation:* trip length vs delay amount

-*Strong positive correlation:* planned/working arrival vs actual arrival

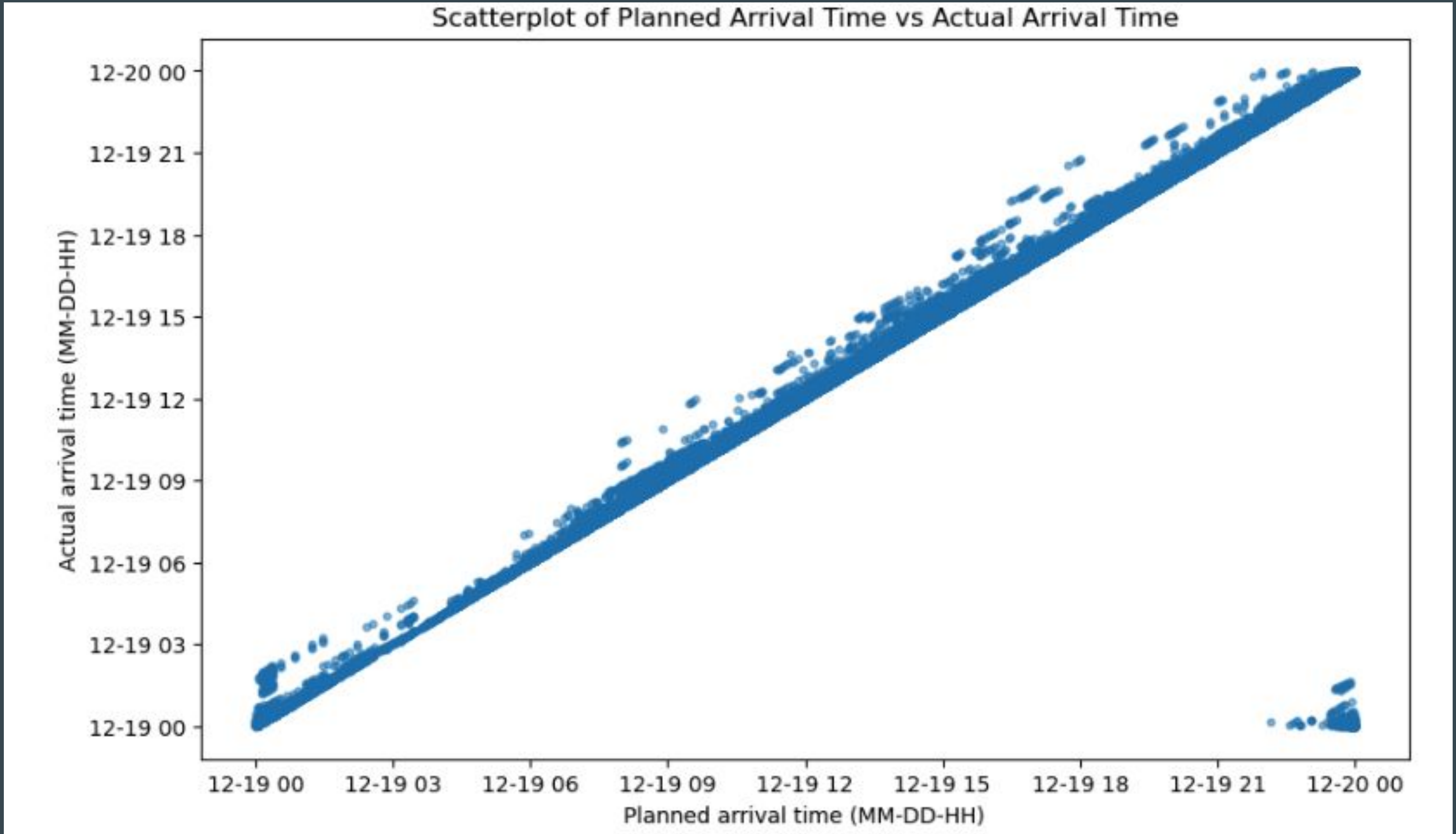


	planned_time_arrival_numeric	working_time_arrival_numeric	actual_arrival_time_numeric	trip_length_numeric	delay_amount_numeric
planned_time_arrival_numeric	1	1	0.99	0.2	-0.042
working_time_arrival_numeric	1	1	0.99	0.2	-0.036
actual_arrival_time_numeric	0.99	0.99	1	0.18	0.11
trip_length_numeric	0.2	0.2	0.18	1	-0.023
delay_amount_numeric	-0.042	-0.036	0.11	-0.023	1

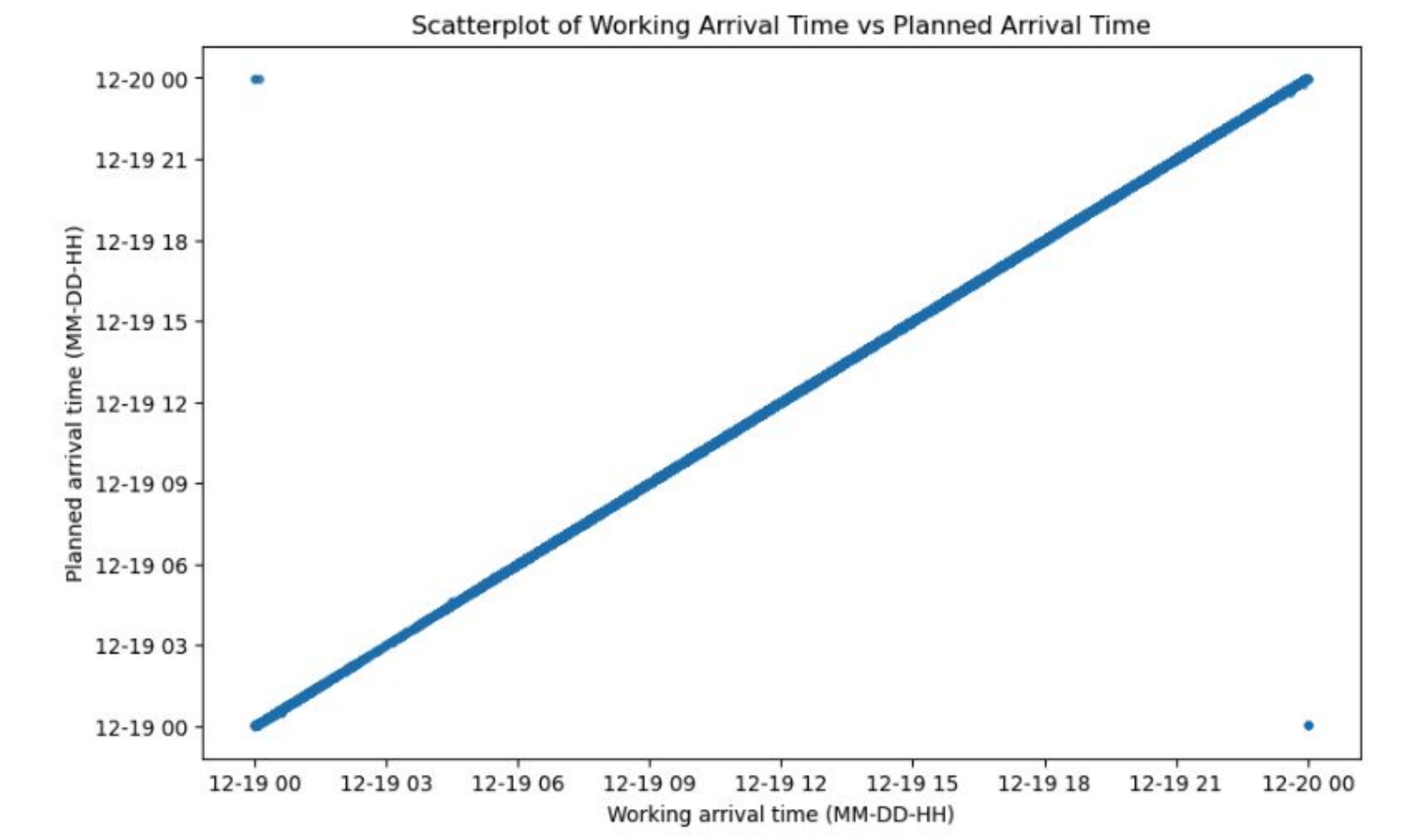
What is the relationship between working arrival time and actual arrival time?



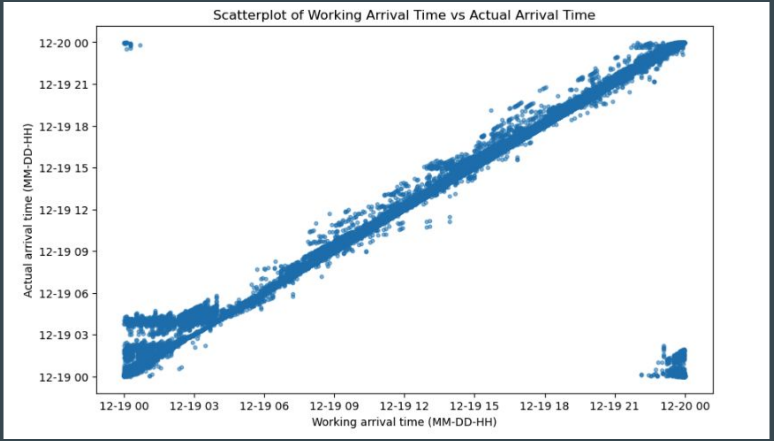
What is the relationship between planned arrival time and actual arrival time?

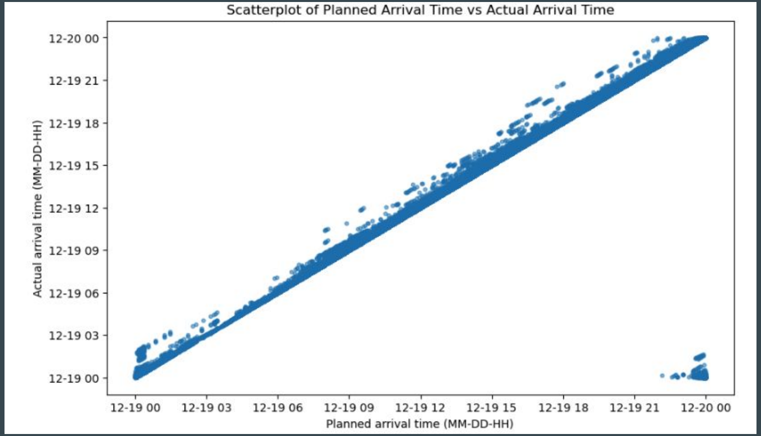


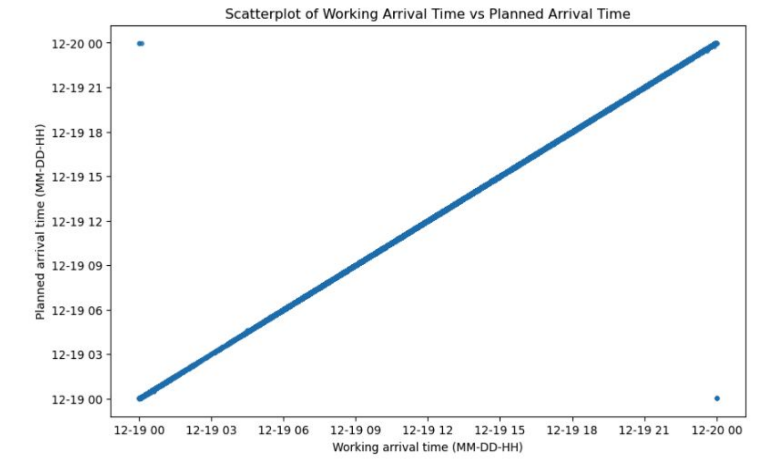
What is the relationship between working arrival time and planned arrival time? How often do they differ?



Overview of time relationships







Conclusions

- Based on data, UK rail system is operating at fairly timely schedule
- Planned vs working times appear to rarely deviate; their current timetables seem to be working well
- Certain factors do affect delays, such as overall trip length
- Considerations on privatization of rail system

Thank you!

Appendix

Project Github: https://github.com/jsnabes/GC_Final_Project

Office of Rail & Road: https://dataportal.orr.gov.uk/statistics/usage/passenger-rail-usage/

UK National Rail maps: https://www.nationalrail.co.uk/travel-information/maps-of-the-national-rail-network/