W205 Final Project: Data Driven

Heather Feinstein
Cynthia Hu
Collin Reinking
Charlotte Swavola

Goals

Empowering taxi drivers with street smarts from data

- How do I optimize my time driving for downtime, but maximize my fares per ride?
- Should I drive back to the area of high demand or wait for a pickup in my current location?
- Is there a strategy to maximise gratuities (tips)?
- Are fewer long-distance trip or more frequent short trips more profitable?
- How should I plan my week to maximize earnings?

Our Roadmap

Phase I: Monthly Dashboard

- Web-based dashboard
- Monthly updated analysis and weekly updated forecast
- Mainly descriptive analysis and simple ML forecast

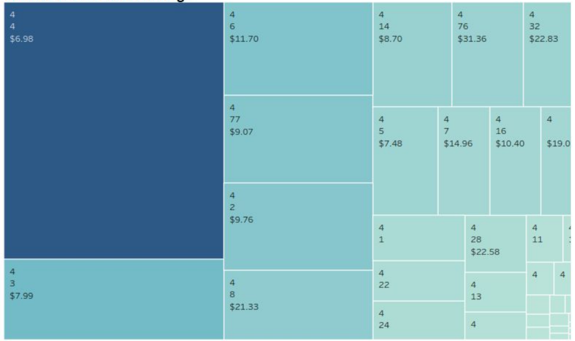
Phase II: Daily or Near-Real-time Application

- Daily or Near-Real-time update
- Advanced ML Fare forecast with improved accuracy
- Integrate both traffic and weather data

Product - Dashboard I

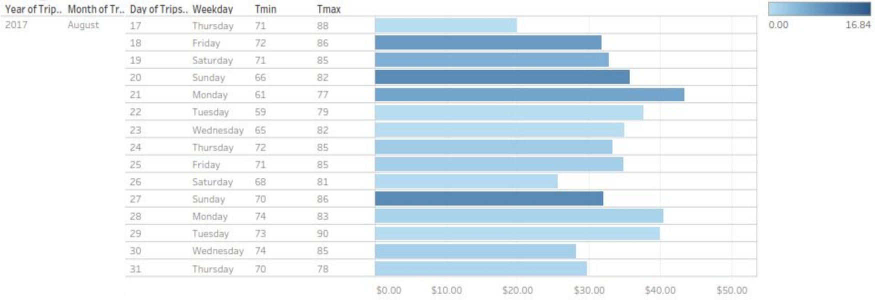
Where are people going from this location?

Current Location:
District 4- Downtown Chicago



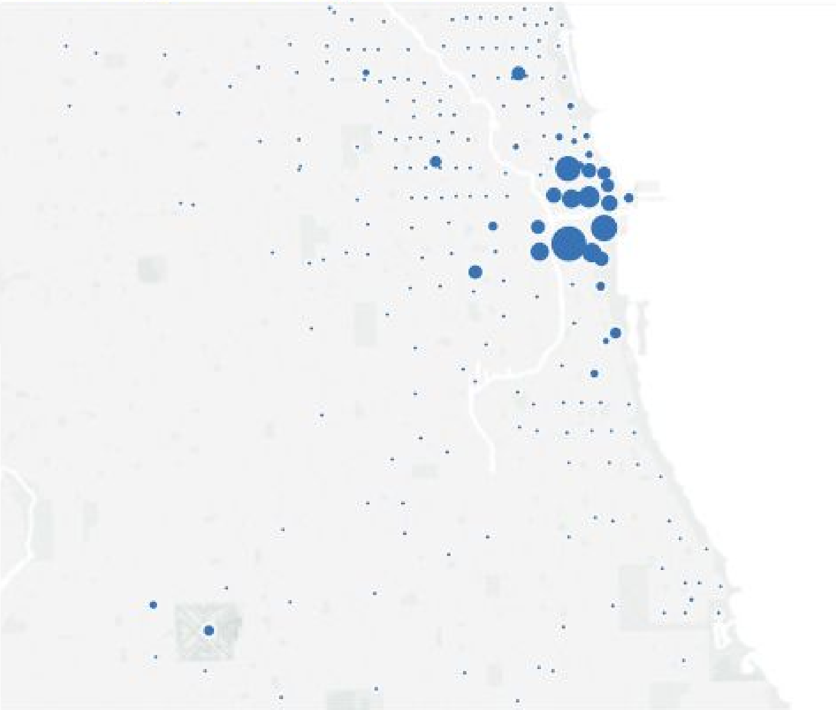
Your Fare (and weather) Forecast

August 2017

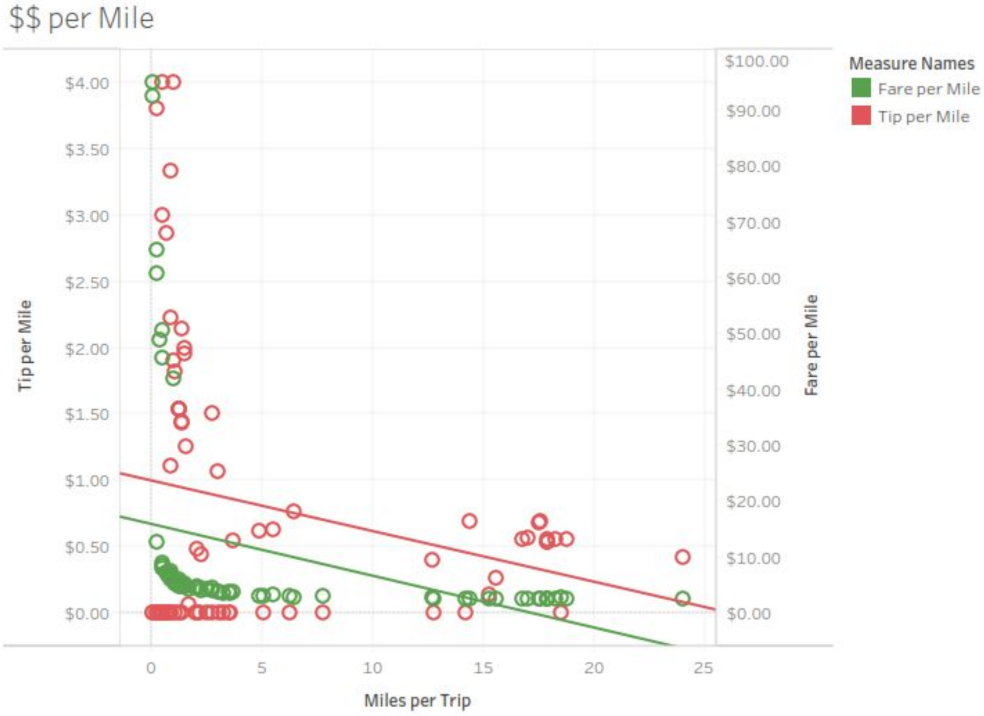


Year of Trip..	Month of Tr..	Day of Trips..	Weekday	Tmin	Tmax
2017	August	17	Thursday	71	88
		18	Friday	72	86
		19	Saturday	71	85
		20	Sunday	66	82
		21	Monday	61	77
		22	Tuesday	59	79
		23	Wednesday	65	82
		24	Thursday	72	85
		25	Friday	71	85
		26	Saturday	68	81
		27	Sunday	70	86
		28	Monday	74	83
		29	Tuesday	73	90
		30	Wednesday	74	85
		31	Thursday	70	78

Chicago Hotspots



Product - Dashboard II (Tableau)



Challenges and Limitations

Challenges:

Correlating consumer demand to pick up data - only one side of the story

Integrating and enabling real-time traffic data and map

Inferring driver down time from available data

Understanding what questions are most critical to drivers

Differentiating ourselves in a crowded marketplace

Dirty data

Limitations:

Historical analysis space requirements

Limit data timeline- more current data offers more relevant information, as external factors (e.g. introducing UberPool) may skew "demand" trends from year to year

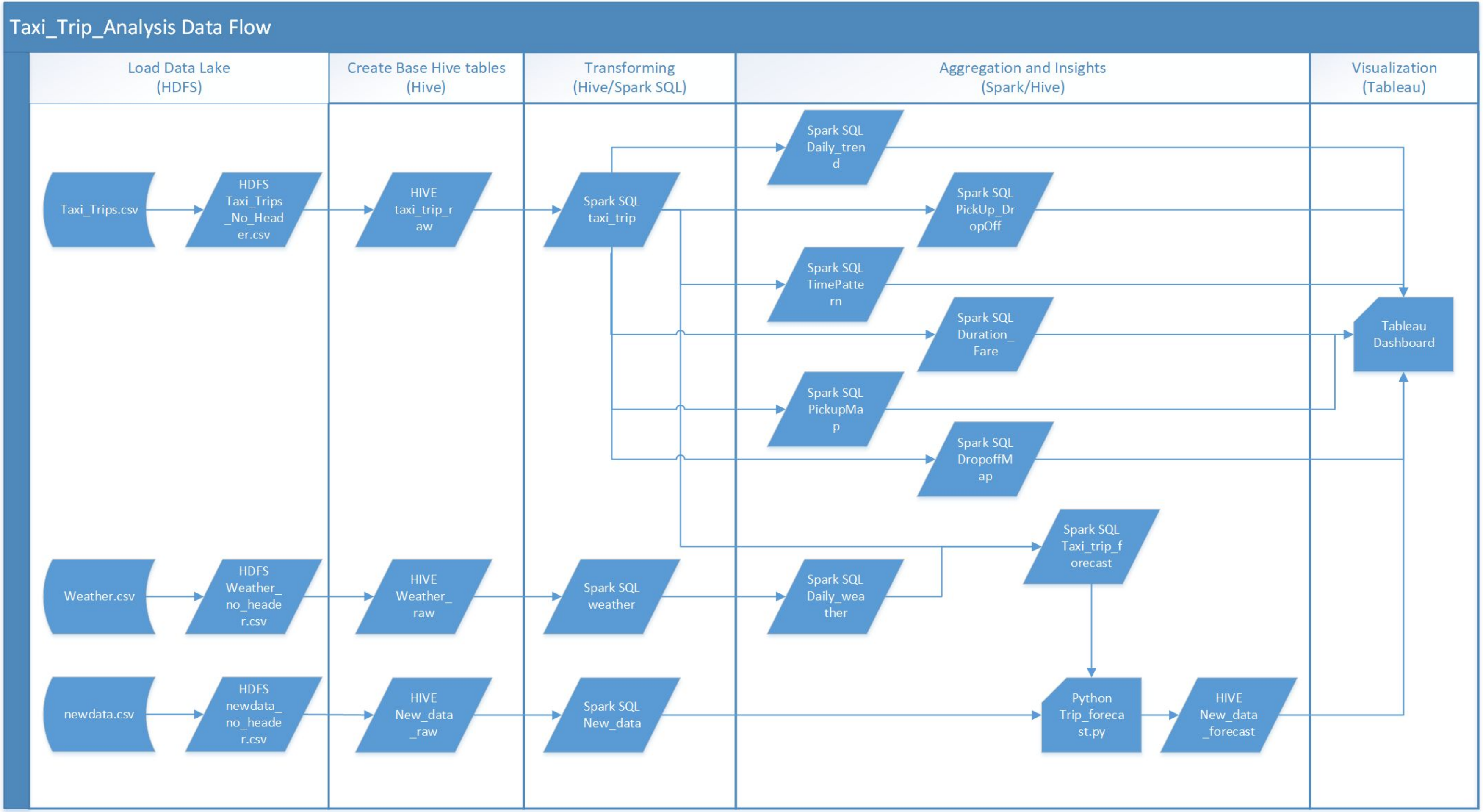
Real-time processing at scale means 6-hour delay

Self fulfilling prophecy- need external data for future forecasting

Storage Requirement

- Challenge: Gathering and transforming the data needs the power of hive, but forecasting and visualization need maneuverable finesse.
- Solutions:
    - Base tables- Hive
        - Data partially filtered from request
    - Table transformations and aggregation- SparkSQL
    - Forecasting- Python code within hive structure
    - Visualization- Tableau! Exciting, informative dashboard, where each graphic answers a question

Architecture



Implementation

1. Set up EC2 Instance with 200+GB EBS volume attached
2. Run load_data.sh to load data and put them into HDFS
3. Run hive_trip_ddl.sql to create base Hive tables
4. Run trip_transforming.sql to transform datasets
5. Run trip_aggregation.sql to aggregate datasets for visualization and forecast
6. Set up python correctly and run trip_forecast.py in PySpark
7. Connect tableau to the Hive Server in EC2 instance
8. Create and refresh charts in tableau

Refer to Implementation Instructions Document

Appendix

Data sets used

1. Chicago Taxi Trip Data https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew

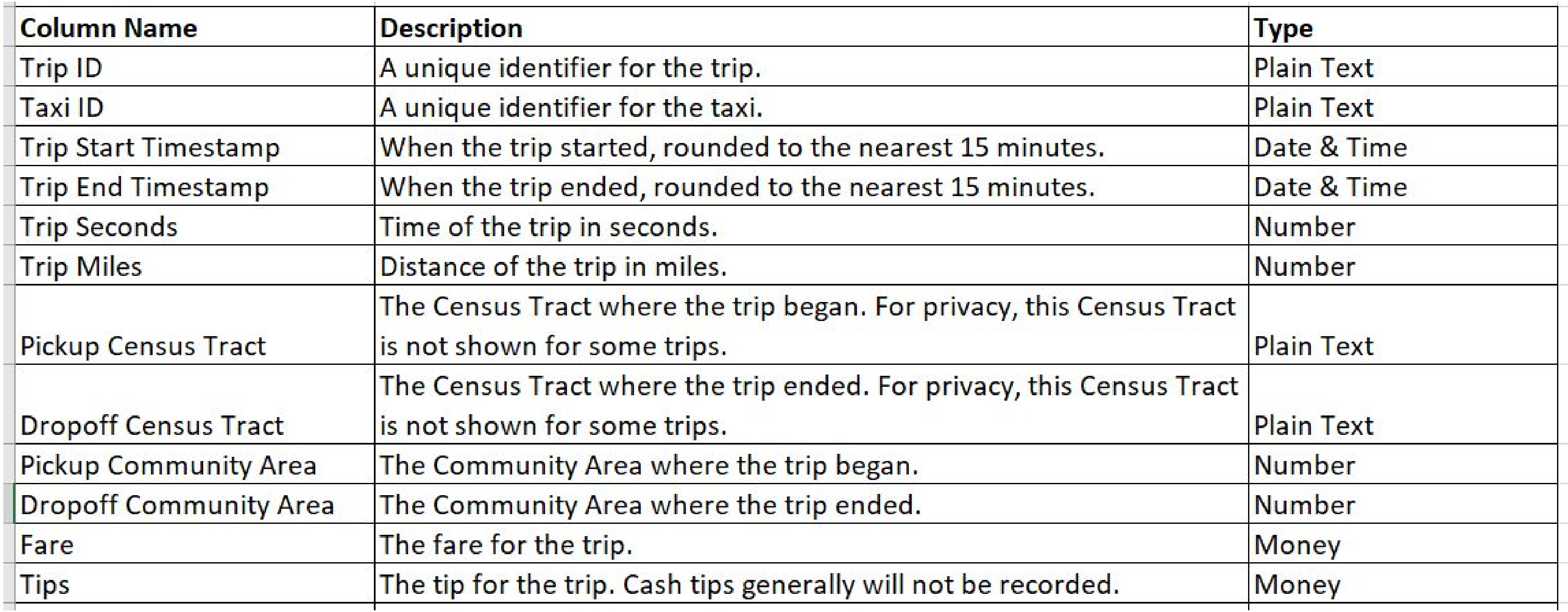
   Time Period: 2013 - Current

   Frequency: Data are updated monthly

   Data Retrieve: Download as CSV, JSON, RDF, TSV or XML; Access through SODA API

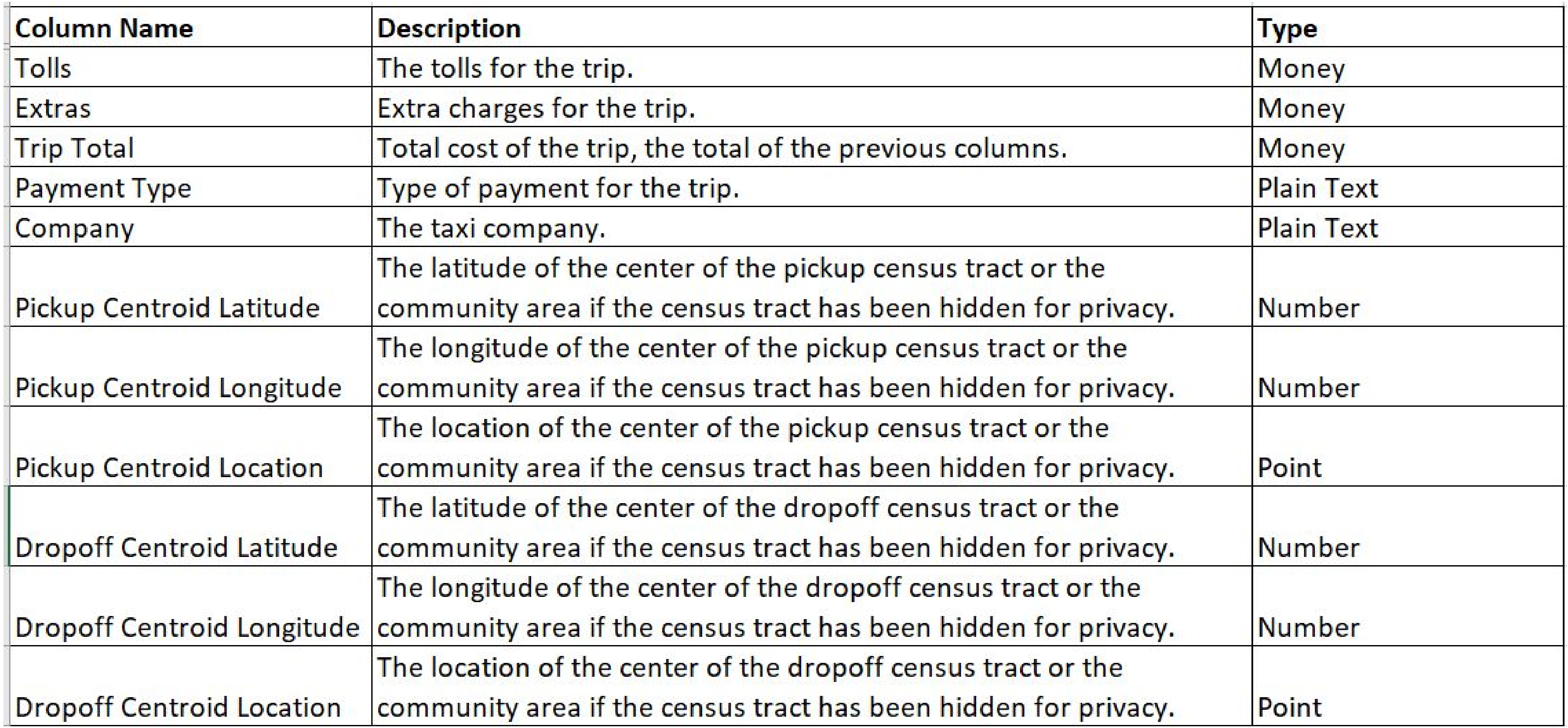
2. Chicago Daily Weather Data
   https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:US170006/detail

Data Dictionary - Taxi Trip I



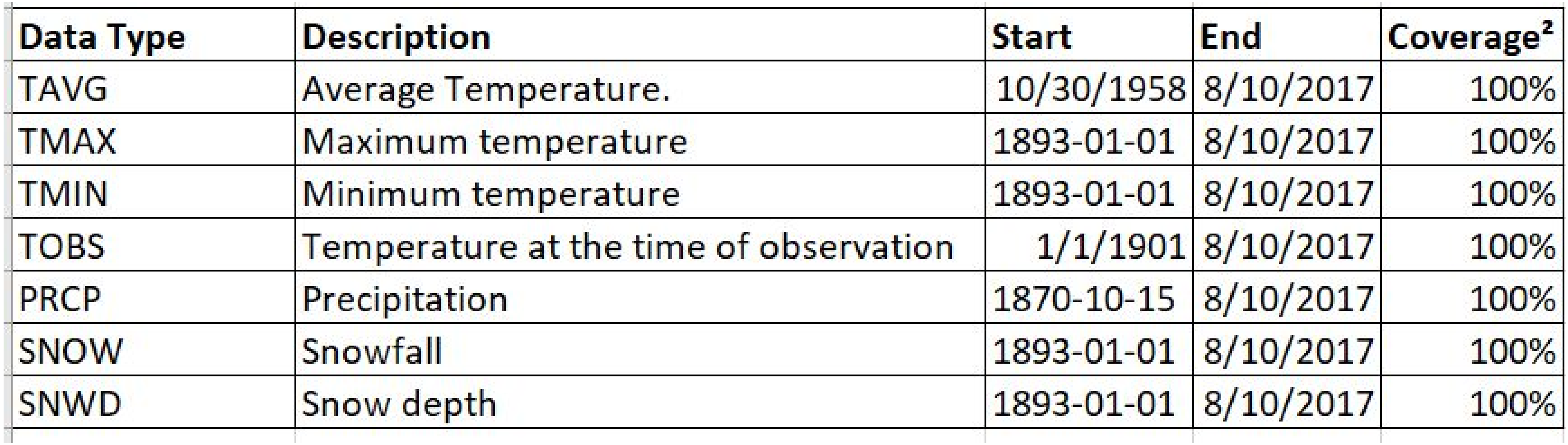
Column Name	Description	Type
Trip ID	A unique identifier for the trip.	Plain Text
Taxi ID	A unique identifier for the taxi.	Plain Text
Trip Start Timestamp	When the trip started, rounded to the nearest 15 minutes.	Date & Time
Trip End Timestamp	When the trip ended, rounded to the nearest 15 minutes.	Date & Time
Trip Seconds	Time of the trip in seconds.	Number
Trip Miles	Distance of the trip in miles.	Number
Pickup Census Tract	The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips.	Plain Text
Dropoff Census Tract	The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips.	Plain Text
Pickup Community Area	The Community Area where the trip began.	Number
Dropoff Community Area	The Community Area where the trip ended.	Number
Fare	The fare for the trip.	Money
Tips	The tip for the trip. Cash tips generally will not be recorded.	Money

Data Dictionary - Taxi Trip II



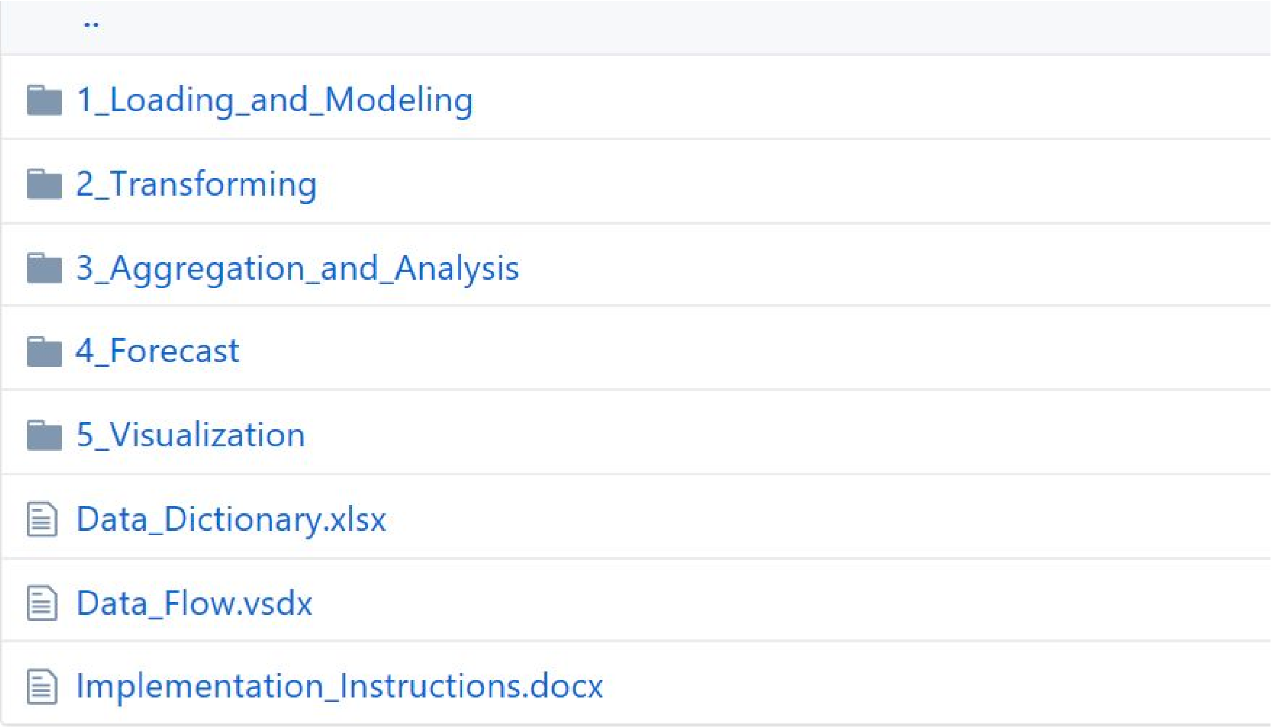
Column Name	Description	Type
Tolls	The tolls for the trip.	Money
Extras	Extra charges for the trip.	Money
Trip Total	Total cost of the trip, the total of the previous columns.	Money
Payment Type	Type of payment for the trip.	Plain Text
Company	The taxi company.	Plain Text
Pickup Centroid Latitude	The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Number
Pickup Centroid Longitude	The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Number
Pickup Centroid Location	The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.	Point
Dropoff Centroid Latitude	The latitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Number
Dropoff Centroid Longitude	The longitude of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Number
Dropoff Centroid Location	The location of the center of the dropoff census tract or the community area if the census tract has been hidden for privacy.	Point

Data Dictionary - Weather



Data Type	Description	Start	End	Coverage[2]
TAVG	Average Temperature.	10/30/1958	8/10/2017	100%
TMAX	Maximum temperature	1893-01-01	8/10/2017	100%
TMIN	Minimum temperature	1893-01-01	8/10/2017	100%
TOBS	Temperature at the time of observation	1/1/1901	8/10/2017	100%
PRCP	Precipitation	1870-10-15	8/10/2017	100%
SNOW	Snowfall	1893-01-01	8/10/2017	100%
SNWD	Snow depth	1893-01-01	8/10/2017	100%

Submission Files



..

1_Loading_and_Modeling

2_Transforming

3_Aggregation_and_Analysis

4_Forecast

5_Visualization

Data_Dictionary.xlsx

Data_Flow.vsdx

Implementation_Instructions.docx

Q & A