ADSP 31014 Statistical Models for Data Science

Course Project Part 1

Overview of Data Set

The data set consists of 13.8M Divvy bike trips and weather information in the City of Chicago from year 2013 to 2017. It was obtained from Kaggle website https://www.kaggle.com/datasets/yingwurenjian/chicago-divvy-bicycle-sharing-data/data (that Kaggle dataset is originally from Divvy Data: https://www.divvybikes.com/system-data and Weather Data: https://www.wunderground.com/). Most of the Divvy trips in the 5 years were represented. Each trip consists of information about its start and end time, pick-up and drop-off location, along with trip duration, user type, gender, weather information (temperature, windchill, dewpoint, humidity, pressure, visibility, wind speed, precipitation, weather condition), and docking station capacity and trip ID. For privacy and ease of calculation, times are around to the nearest minutes.

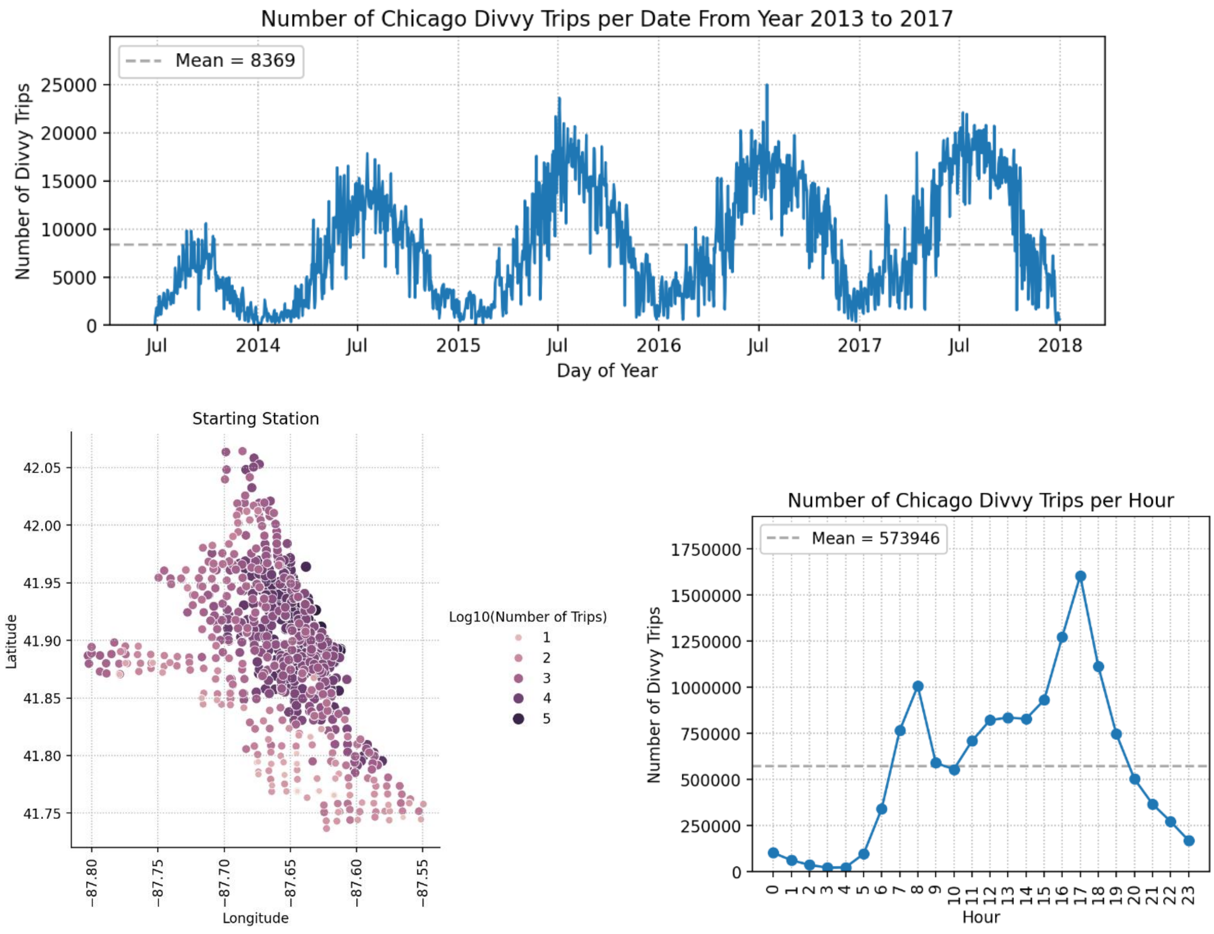

Figure 1: The number of Divvy trips per date (top), starting station (low left), and hour (low right)

Data Table Schema

The data set consists of 13,774,715 rows and 27 columns. Each row represents one Divvy trip.

Column Name	Description	Data Type	Example Value	Notes
trip_id	Unique ID per trip	integer	4192	• 13,774,654 unique trip ID • 61 rows of duplicated data
usertype	Type of rider	string	Subscriber	• Subscriber (72.7%) – Mostly • Customer (27%) • Dependent (0.3%)
gender	Gender of rider	string	Male	• 3,756,932 rows of missing data • Male (75%) – Mostly • Female (25%)
starttime	Start date and time	string	2013-06-27 12:15:00	• Had daylight saving (only in 2017) • 11 rows of ending time before starting time
stoptime	End date and time	string	2013-06-27 12:16:00	
tripduration	Duration of trip in seconds	integer	60	• Range from 1min to 24 hours • 302,597 rows of data larger than 1 hour
from_station_id	Unique ID for starting station	integer	28	• 586 rows of unique station IDs • 663 rows of unique station names • 145 rows of station names with duplicated station IDs (typos, etc.)
from_station_name	Name of starting station	string	Larrabee St & Menomonee St	
latitude_start	Latitude of starting station	decimal	41.91468	• 1,153 rows of missing data for each • Millennium Park, Navy Pier, downtown CBD area have most popular starting and ending stations
longitude_start	Longitude of starting station	decimal	-87.64332	
dpcapacity_start	Docking capacity at starting station	decimal	15.0	
to_station_id	Unique ID for ending station	integer	28	• Exactly same as “from_station_id” and “from_station_name” • Each station will serve both as starting and ending stations
to_station_name	Name of ending station	string	Larrabee St & Menomonee St	
latitude_end	Latitude of ending station	decimal	41.91468	• 1,180 rows of missing data for each • Starting and ending latitude/longitude are internally consistent • Docking capacity range from 0 to 55
longitude_end	Longitude of ending station	decimal	-87.64332	
dpcapacity_end	Docking capacity at ending station	decimal	15.0	
temperature	Temperature of trip (°F)	decimal	87.1	• 858 rows of anomalous temperature data (-9999 °F) • 11,837,251 rows of anomalous windchill data (-999 °F) • 920 rows of anomalous dewpoint data (-9999 °F) • 920 rows of missing humidity data • 4,442 rows of anomalous pressure data (-9999 inHg)
windchill	Windchill of trip (°F)	decimal	-999.0	
dewpoint	Dew point of trip (°F)	decimal	69.1	
humidity	Humidity of trip (%)	decimal	55.0	
pressure	Atmospheric pressure of trip (inHg)	decimal	29.75	

visibility	Visibility of trip (miles)	decimal	10.0	• 2,358 rows of anomalous visibility data (-9999 miles) • 4,253 rows of anomalous wind speed data (-9999 mph) • 12,833,368 rows of anomalous precipitation data (-9999 inches)
wind_speed	Wind speed of trip (mph)	decimal	13.8	
precipitation	Precipitation of trip (inches)	decimal	-9999.0	
events	Weather events of trip	string	mostlycloudy	• Mostly/Partly/Scattered Cloudy (89%) - Mostly • Clear (5.5%) • Rainy (3%) • Stormy (1%) • Snowy (1%) • Hazy (0.5%)
rain	Binary indicator of rain occurrence during trip	integer	0	
conditions	Overall weather conditions of trip	string	Mostly Cloudy	

Data Cleaning and Processing

The two main data quality issues are missing gender information and outliers in trip duration, windchill and precipitation. Missing gender data seems not to be missing at random, with seasonality trends in it.

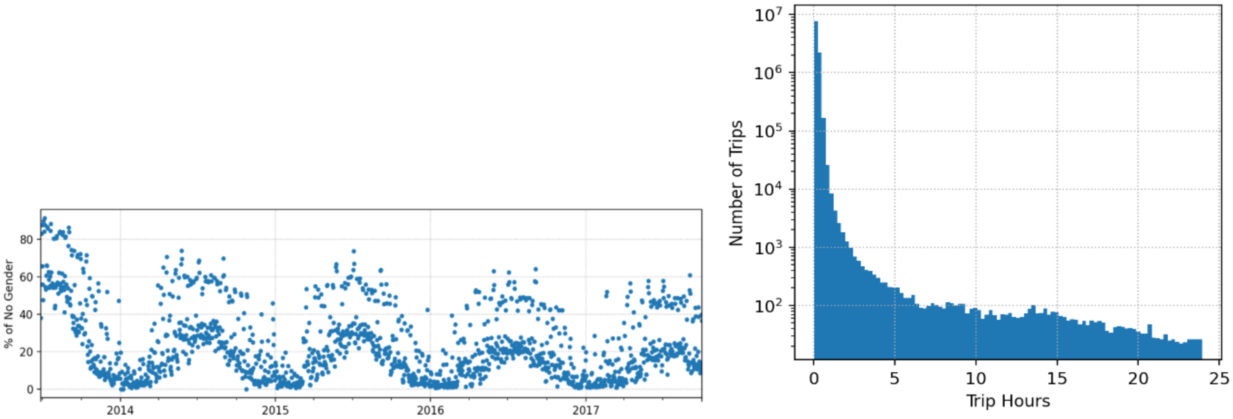

Figure 2: Missing gender data (left) and outliers of trip duration (right)

Consequently, the data set is cleaned by removing rows of the following kind:

- Row with any missing and duplicated data
- Starttime > stoptime
- Tripduration outside 1 second to 1 hour
- Temperature outside 30-90 °F, windchill outside 0-45 °F, dewpoint outside 0-80 °F, pressure outside of 29-31 inHg, visibility outside 0-10 miles, wind speed outside 0-43 mph, precipitation outside 0-0.1 inches

Moreover, 2 typos in Divvy station name are fixed (some typos remained). Weather conditions not of "Cloudy" and "Clear" are re-categorized as "Others". The resulting cleaned data set has 2,355,134 rows.