Exercise 6.1

Daniel Hurrle

# Data Source

**Description:**

The data source was published by Ryan Cummings on Kaggle.com and contains data about the mobility service Citi Bike from its launch until October 2013.

The data is published on trip level containing the date and time, the location of the start and end trip, and some general demographic data (if available) such as gender, and birth year.

The total count of rows is 50,000.

**Motivation:**

I’ve chosen this data source for analysis as I’ve been working in the mobility space in my previous career and am interested in mobility patterns, especially in the shared space.

**Source of Data:**

https://www.kaggle.com/datasets/ryanmcummings/citi-bike-data

# Data Profile

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Example** | **Data Type** | **Description** |
| **trip\_id** | **LnQzQk** | **object** | **Trip ID** |
| **bike\_id** | **16013** | **int64** | **Bike ID** |
| **weekday** | **Mon** | **object** | **Day of the week** |
| **start\_hour** | **18** | **int64** | **Hour of Start Trip** |
| **start\_time** | **09.09.13 18:18** | **object** | **Start time of Trip** |
| **start\_station\_id** | **523** | **int64** | **Station ID (Trip Start)** |
| **start\_station\_name** | **W 38 St & 8 Ave** | **object** | **Station Name (Trip Start)** |
| **start\_station\_latitude** | **40,75466591** | **float64** | **Station Latitude (Trip Start)** |
| **start\_station\_longitude** | **-73,99138152** | **float64** | **Station Longitude (Trip Start)** |
| **end\_time** | **09.09.13 18:35** | **object** | **End time of Trip** |
| **end\_station\_id** | **334** | **int64** | **Station ID (Trip End)** |
| **end\_station\_name** | **W 20 St & 7 Ave** | **object** | **Station Name (Trip End)** |
| **end\_station\_latitude** | **40,74238787** | **float64** | **Station Latitude (Trip End)** |
| **end\_station\_longitude** | **-73,99726235** | **float64** | **Station Longitude (Trip End)** |
| **trip\_duration** | **993** | **int64** | **Duration of trip in minutes** |
| **subscriber** | **Subscriber** | **object** | **Subscriber vs Non-Subscriber** |
| **birth\_year** | **1968** | **float64** | **Year of Birth of Customer** |
| **gender** | **2** | **int64** | **Gender of Customer** |

**Data types:**

The data contains 5 float64, 6 int64 and 7 object data types and thus is in line with the requirements for this task.

**Data Cleaning:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Change** | **Rows affected** | **% of Total Rows** |
| birth\_year | Removed 'NA' | 6979 | 0,13958 |
| gender | removed '0' | 6981 | 0,13962 |

**Descriptive Statistics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***bike\_id*** | |  | ***start\_hour*** | |
|  |  |  |  |  |
| Mean | 17615,2694 |  | Mean | 14,14524 |
| Standard Error | 7,49264988 |  | Standard Error | 0,021737 |
| Median | 17584 |  | Median | 15 |
| Mode | 16188 |  | Mode | 17 |
| Standard Deviation | 1675,40745 |  | Standard Deviation | 4,8605409 |
| Sample Variance | 2806990,11 |  | Sample Variance | 23,6248578 |
| Kurtosis | -1,1582525 |  | Kurtosis | -0,2698257 |
| Skewness | 0,0230013 |  | Skewness | -0,4539925 |
| Range | 6086 |  | Range | 23 |
| Minimum | 14556 |  | Minimum | 0 |
| Maximum | 20642 |  | Maximum | 23 |
| Sum | 880763468 |  | Sum | 707262 |
| Count | 50000 |  | Count | 50000 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***start\_station\_id*** | |  | ***end\_station\_id*** | |
|  |  |  |  |  |
| Mean | 443,3215 |  | Mean | 442,5397 |
| Standard Error | 1,59458446 |  | Standard Error | 1,5909893 |
| Median | 402 |  | Median | 402 |
| Mode | 459 |  | Mode | 497 |
| Standard Deviation | 356,559925 |  | Standard Deviation | 355,756022 |
| Sample Variance | 127134,98 |  | Sample Variance | 126562,347 |
| Kurtosis | 22,4900773 |  | Kurtosis | 22,2330899 |
| Skewness | 4,48678537 |  | Skewness | 4,4659459 |
| Range | 2930 |  | Range | 2930 |
| Minimum | 72 |  | Minimum | 72 |
| Maximum | 3002 |  | Maximum | 3002 |
| Sum | 22166075 |  | Sum | 22126985 |
| Count | 50000 |  | Count | 50000 |

|  |  |
| --- | --- |
| ***trip\_duration*** | |
|  |  |
| Mean | 838,9829 |
| Standard Error | 2,56550339 |
| Median | 672 |
| Mode | 2697 |
| Standard Deviation | 573,663997 |
| Sample Variance | 329090,382 |
| Kurtosis | 1,45271124 |
| Skewness | 1,30805023 |
| Range | 2637 |
| Minimum | 60 |
| Maximum | 2697 |
| Sum | 41949145 |
| Count | 50000 |
|  |  |
|  |  |

**Limitations and Ethics**

The data presented comes from the customer base of a specific mobility service. Population data thus are not representative for the general population in the geo.

Data can not be verified as it was collected by the company running the service.

Gender data is not clear, as we do not now whether 1 stands for male or female.

# Key Questions

1. Trip characteristics: how long was the longest trips, the shortest trip and what’s the average trip duration?
2. Trip destinations: Are most trips in the city center? Which are the busiest districts?
3. When do most trips happen? Do they follow the rush hour logic? What about weekday / weekend?
4. Who is using the service? How is the age distributed in the customer base?
5. How has the service evolved over time (number of trips) per year?