

Data Exploration Project – 402251

Uber & Weather Dataset

This project was conducted by:

Ghazal Almidana

Introduction:

**Abstract:**

New York city is one of the most famous cities in the world with over 18.000.000 million population. the weather is mostly unstable through the day and it effects many things starting from our mood at the start of the day to the traffic and how the cars and drivers get effected as will from the bad weather. So that will lead us to our research which was to prove our theory which is "How the Weather Effect the Performance of Uber". And the procedure that was followed for this research was by Data Analysis.

**Intro:**

New York city is one of the most famous cities in the world. it faces a lot of trouble with the changing weather and even if it's not expected it's very common to have natural disasters in this city and that would affect all the different aspects of life. The weather dataset collected by NOAA (National Oceanic and Atmospheric Administration) from over 35,000 thousand weather stations all over the world have a daily data for many variables that are related to the weather. And also we have the uber dataset that was collected by uber and it has also variables that are related to the uber performance.

**The Scope:**

This project was conducted to highlight how the Uber performance is effected by the weather through studying the weather changes starting from January of 2009 to June of the year 2015 which was almost the start of uber and of course the weather data was followed to match the uber dataset so we selected a dataset that was in the same time period as the uber dataset so we can merge them and start our search.

This type of project and the studies that went in it could be used for many stakeholders that are either related to the uber and delivery organizations or even other stakeholders that could benefit from the study since the weather is general aspect in the business life and many companies will be affected from the weather.

**Problem Definition:**

When the uber started it needed a way to increase the profit and to spread more through new York and the whole world so when starting it faced a problem with the taxi since it's basically the same business but with different details of the way it works and the whole procedure and then it started increasing by the years but of course with the pollution that is increasing by the day we get in more troubles each day and for that we need more creative ways to expand the performance of the uber drivers and uber fares.

**First The datasets - Exploring:**

The uber details that is collected from the original data without any altering with the documentation from the original dataset collector.

1.The UBER Dataset:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Columns Name | Description | Data Type |
| 1 | Trip\_DATE | The date the trip took place. Given in mm/dd/yyyy, 00:00:00 format | Object |
| 2 | Trip\_ID | what is the trip id for a selected trip during the day. | Object |
| 3 | fare\_amount | is the total fares for a single day. | Float |
| 4 | pickup\_longitude | The longitude of the pickup location. | Float |
| 5 | pickup\_latitude | The latitude of the pickup location. | Float |
| 6 | dropoff\_longitude | The longitude of the drop off location. | Float |
| 7 | dropoff\_latitude | The latitude of the drop off location. | Float |
| 8 | passenger\_count | the total number of passengers for each day. | Object |

The weather details that is collected from the original data without any altering with the documentation from the original dataset collector.

2.The weather Dataset (NOAA):

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | STATION | Station number (WMO/DATSAV3 possibly combined w/WBAN number. | Float |
| 2 | LATITUDE | Given in decimated degrees (Southern Hemisphere values are negative). | Float |
| 3 | LONGITUDE | Given in decimated degrees (Western Hemisphere values are negative). | Float |
| 4 | ELEVATION | Given in meters. | Float |
| 5 | NAME | Name of station/airport/military base. | Object |
| 6 | TEMP | Mean temperature for the day in degrees Fahrenheit to tenths. Missing = 9999.9 | Float |
| 7 | TEMP\_ATTRIBUTES | Number of observations used in calculating mean temperature. | Object |
| 8 | DEWP | Mean dew point for the day in degrees Fahrenheit to tenths. Missing = 9999.9 | Float |
| 9 | DEWP\_ATTRIBUTES | Number of observations used in calculating mean dew point. | Object |
| 10 | SLP | Mean sea level pressure for the day in millibars to tenths. Missing = 9999.9 | Float |
| 11 | SLP\_ATTRIBUTES | Number of observations used in calculating mean sea level pressure. | Object |
| 12 | STP | Mean station pressure for the day in millibars to tenths. Missing = 9999.9 | Float |
| 13 | STP\_ATTRIBUTES | Number of observations used in calculating mean station pressure. | Object |
| 14 | VISIB | Mean visibility for the day in miles to tenths. Missing = 999.9 | Float |
| 15 | VISIB\_ATTRIBUTES | Number of observations used in calculating mean visibility. | Object |
| 16 | WDSP | Mean wind speed for the day in knots to tenths. Missing = 999.9 | Float |
| 17 | WDSP\_ATTRIBUTES | Number of observations used in calculating mean wind speed. | Object |
| 18 | MXSPD | Maximum sustained wind speed reported for the day in knots to tenths. Missing =  999. | Float |
| 19 | GUST | Maximum wind gust reported for the day in knots to tenths. Missing = 999.9 | Float |
| 20 | MAX | Maximum temperature reported during the day in Fahrenheit to tenths. Missing = 9999.9 | Float |
| 21 | MAX\_ATTRIBUTES | Blank indicates maximum temperature was taken from the explicit maximum  temperature report and not from the 'hourly' data. | Object |
| 22 | MIN | Minimum temperature reported during the day in Fahrenheit to tenths. Missing = 9999.9 | Float |
| 23 | MIN\_ATTRIBUTES | Blank indicates minimum temperature was taken from the explicit minimum  temperature report and not from the 'hourly' data.  \* indicates minimum temperature was derived from the hourly data  (i.e. highest hourly or synoptic-reported temperature). | Object |
| 24 | PRCP | Total precipitation (rain and/or melted snow) reported during the day in inches  and hundredths; will usually not end with the midnight observation (i.e. may include  latter part of the previous day). “0” indicates no measurable precipitation (includes a trace).  Missing = 99.99 | Float |
| 25 | PRCP\_ATTRIBUTES | A = 1 report of 6-hour precipitation amount.  B = Summation of 2 reports of 6-hour precipitation amount.  C = Summation of 3 reports of 6-hour precipitation amount.  D = Summation of 4 reports of 6-hour precipitation amount.  E = 1 report of 12-hour precipitation amount.  F = Summation of 2 reports of 12-hour precipitation amount.  G = 1 report of 24-hour precipitation amount.  H = Station reported '0' as the amount for the day (eg, from 6-hour reports),but also reported at least one occurrence of precipitation in hourly observations.  This could indicate a trace occurred, but should be considered as incomplete data for the day.  I = Station did not report any precipitation data for the day and did not report any  occurrences of precipitation in its hourly observations. It's still possible that  precipitation occurred but was not reported. | Object |
| 26 | FRSHTT | Indicators (1 = yes, 0 = no/not reported) for the occurrence during the day of:  Fog ('F' - 1st digit).  Rain or Drizzle ('R' - 2nd digit).  Snow or Ice Pellets ('S' - 3rd digit).  Hail ('H' - 4th digit).  Thunder ('T' - 5th digit).  Tornado or Funnel Cloud ('T' - 6th digit). | Integer |
| 27 | Date | Given in mm/dd/yyyy format | object |

First the two datasets was merged together using merge function from pandas library so we can now proceed to the next steps with our full data.

We started exploring our data with different functions and visual plots:

Like numpy and missingno to check whether their was duplicated, missing values or even white spaces and all that was performed and can be found in the ZIP file under the name "Cleaning.py" with the documentation for each exploration step.

**Second Part: Cleaning the two Datasets.**

This part was started after first merging the two data sets together then we figured that it was mentioned in the documentation that each missing value is replaced by either 999.9 or 9999.9 and that would increase the mean and distribution and any other values to a drastically Large value and that would give wrong visualizations and insights so we begin the cleaning process by:

Replacing the missing values with the mean or median with what fits the data best and then removed any duplicates and lastly made new date column with the format mm/dd/yyyy without the hours and other cleaning steps which is all specified in the file under the name "Cleaning.py".

And here we have the merged clean Dataset variables:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Columns Name | Description | Data Type |
| 1 | Trip\_DATE | The date the trip took place. | Object |
| 2 | Trip\_ID | what is the trip id for a selected trip during the day. | Integer |
| 3 | fare\_amount | is the total fares for a single day. | Float |
| 4 | pickup\_longitude | The longitude of the pickup location. | Float |
| 5 | pickup\_latitude | The latitude of the pickup location. | Float |
| 6 | dropoff\_longitude | The longitude of the drop off location. | Float |
| 7 | dropoff\_latitude | The latitude of the drop off location. | Float |
| 8 | passenger\_count | the total number of passengers for each day. | Integer |
| 9 | STATION | Station number (WMO/DATSAV3 possibly combined w/WBAN number. | Float |
| 10 | LATITUDE | Given in decimated degrees (Southern Hemisphere values are negative). | Float |
| 11 | LONGITUDE | Given in decimated degrees (Western Hemisphere values are negative). | Float |
| 12 | ELEVATION | Given in meters. | Float |
| 13 | NAME | Name of station/airport/military base. | Object |
| 14 | TEMP | Mean temperature for the day in degrees Fahrenheit to tenths. Missing = 9999.9 | Float |
| 15 | TEMP\_ATTRIBUTES | Number of observations used in calculating mean temperature. | Integer |
| 16 | DEWP | Mean dew point for the day in degrees Fahrenheit to tenths. Missing = 9999.9 | Float |
| 17 | DEWP\_ATTRIBUTES | Number of observations used in calculating mean dew point. | Integer |
| 18 | SLP | Mean sea level pressure for the day in millibars to tenths. Missing = 9999.9 | Float |
| 19 | SLP\_ATTRIBUTES | Number of observations used in calculating mean sea level pressure. | Integer |
| 20 | STP | Mean station pressure for the day in millibars to tenths. Missing = 9999.9 | Float |
| 21 | STP\_ATTRIBUTES | Number of observations used in calculating mean station pressure. | Integer |
| 22 | VISIB | Mean visibility for the day in miles to tenths. Missing = 999.9 | Float |
| 23 | VISIB\_ATTRIBUTES | Number of observations used in calculating mean visibility. | Integer |
| 24 | WDSP | Mean wind speed for the day in knots to tenths. Missing = 999.9 | Float |
| 25 | WDSP\_ATTRIBUTES | Number of observations used in calculating mean wind speed. | Integer |
| 26 | MXSPD | Maximum sustained wind speed reported for the day in knots to tenths. Missing =  999. | Float |
| 27 | GUST | Maximum wind gust reported for the day in knots to tenths. Missing = 999.9 | Float |
| 28 | MAX | Maximum temperature reported during the day in Fahrenheit to tenths. Missing = 9999.9 | Float |
| 29 | MAX\_ATTRIBUTES | Blank indicates maximum temperature was taken from the explicit maximum  temperature report and not from the 'hourly' data.  \* indicates maximum temperature was derived from the hourly data  (i.e. highest hourly or synoptic-reported temperature). | Object |
| 30 | MIN | Minimum temperature reported during the day in Fahrenheit to tenths. Missing = 9999.9 | Float |
| 31 | MIN\_ATTRIBUTES | Blank indicates minimum temperature was taken from the explicit minimum  temperature report and not from the 'hourly' data.  \* indicates minimum temperature was derived from the hourly data  (i.e. highest hourly or synoptic-reported temperature). | Object |
| 32 | PRCP | Total precipitation (rain and/or melted snow) reported during the day in inches  and hundredths; will usually not end with the midnight observation (i.e. may include  latter part of the previous day). “0” indicates no measurable precipitation (includes a trace).  Missing = 99.99 | Float |
| 33 | PRCP\_ATTRIBUTES | A = 1 report of 6-hour precipitation amount.  B = Summation of 2 reports of 6-hour precipitation amount.  C = Summation of 3 reports of 6-hour precipitation amount.  D = Summation of 4 reports of 6-hour precipitation amount.  E = 1 report of 12-hour precipitation amount.  F = Summation of 2 reports of 12-hour precipitation amount.  G = 1 report of 24-hour precipitation amount.  H = Station reported '0' as the amount for the day (eg, from 6-hour reports),but also reported at least one occurrence of precipitation in hourly observations.  This could indicate a trace occurred, but should be considered as incomplete data for the day.  I = Station did not report any precipitation data for the day and did not report any  occurrences of precipitation in its hourly observations. It's still possible that  precipitation occurred but was not reported. | Object |
| 34 | FRSHTT | Indicators (1 = yes, 0 = no/not reported) for the occurrence during the day of:  Fog ('F' - 1st digit).  Rain or Drizzle ('R' - 2nd digit).  Snow or Ice Pellets ('S' - 3rd digit).  Hail ('H' - 4th digit).  Thunder ('T' - 5th digit).  Tornado or Funnel Cloud ('T' - 6th digit). | Integer |
| 35 | Date | Given in mm/dd/yyyy format | DateTime |

**Last Step – Visualizing The Dataset:**

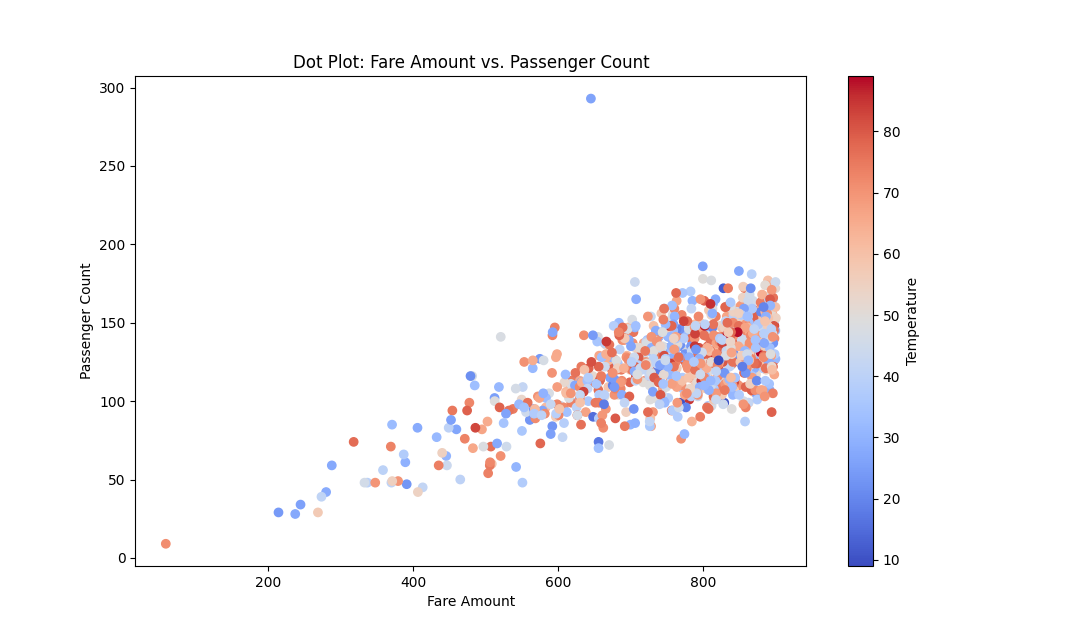
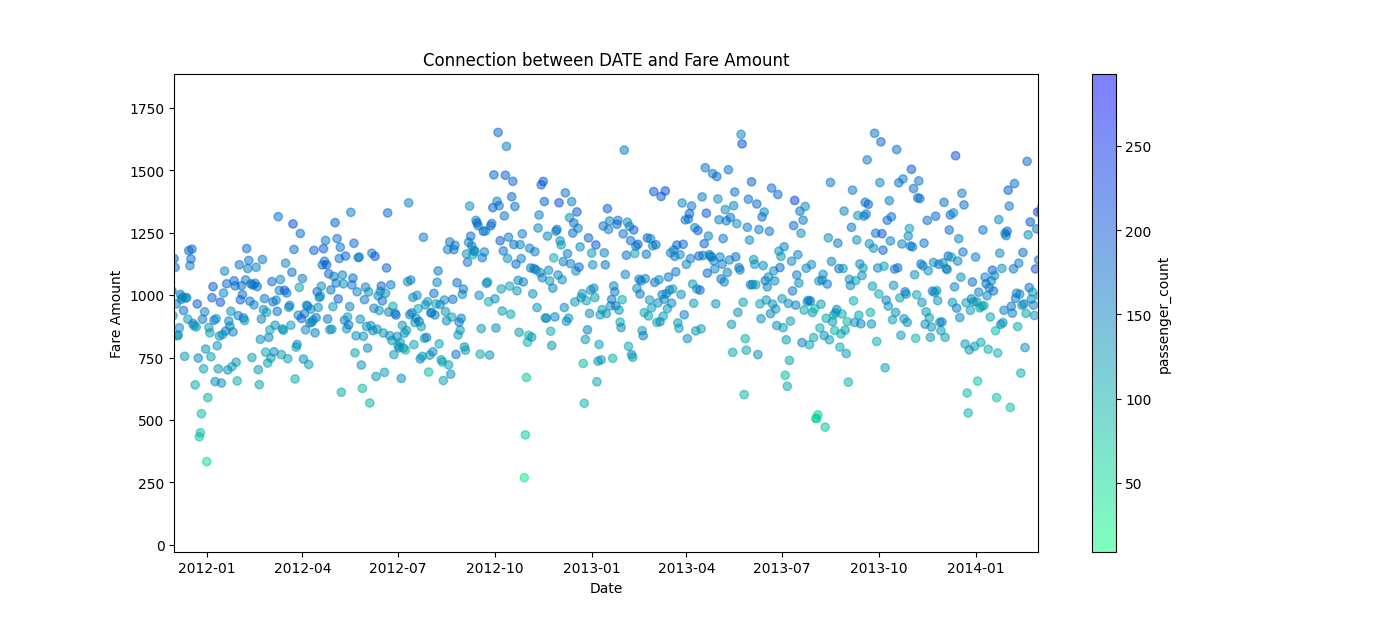
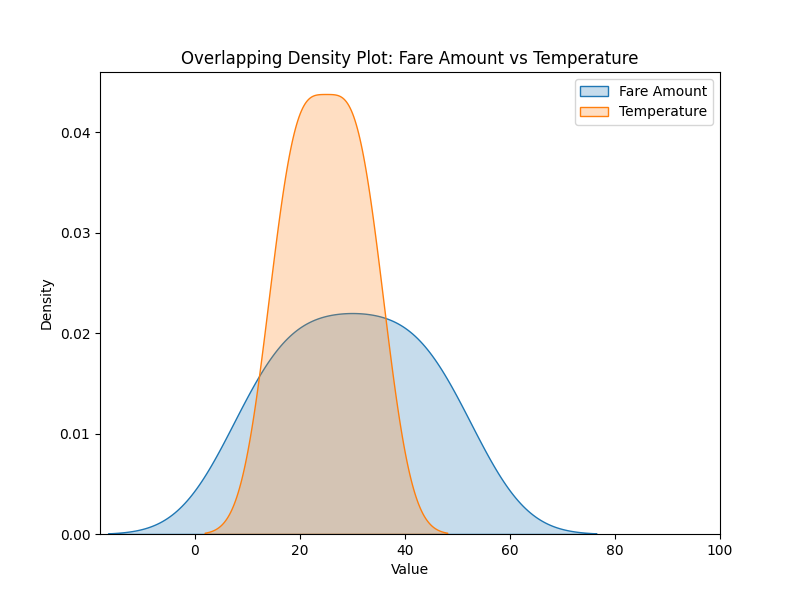
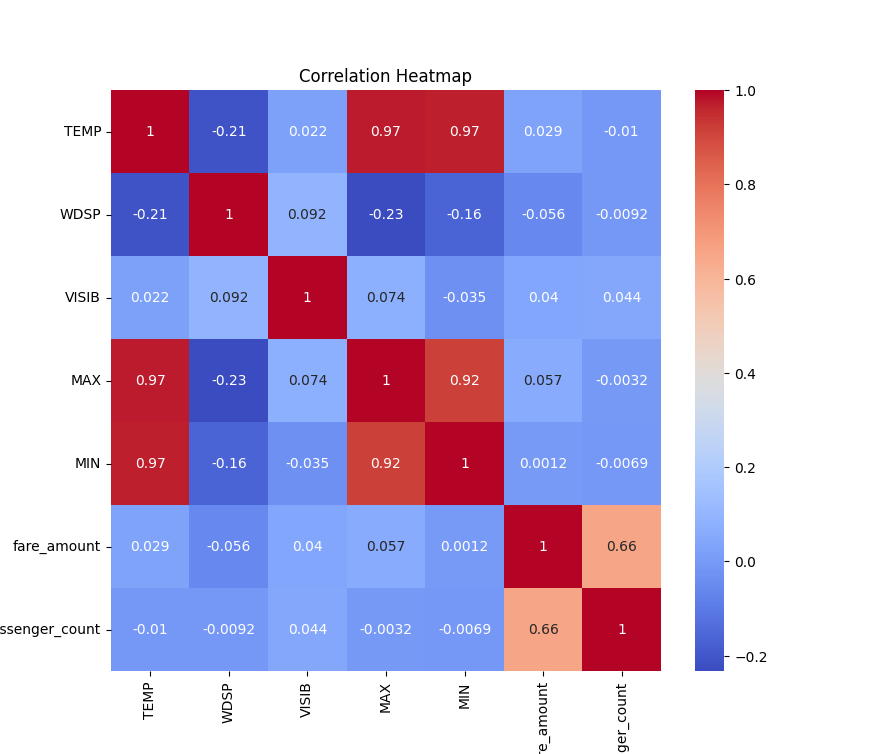
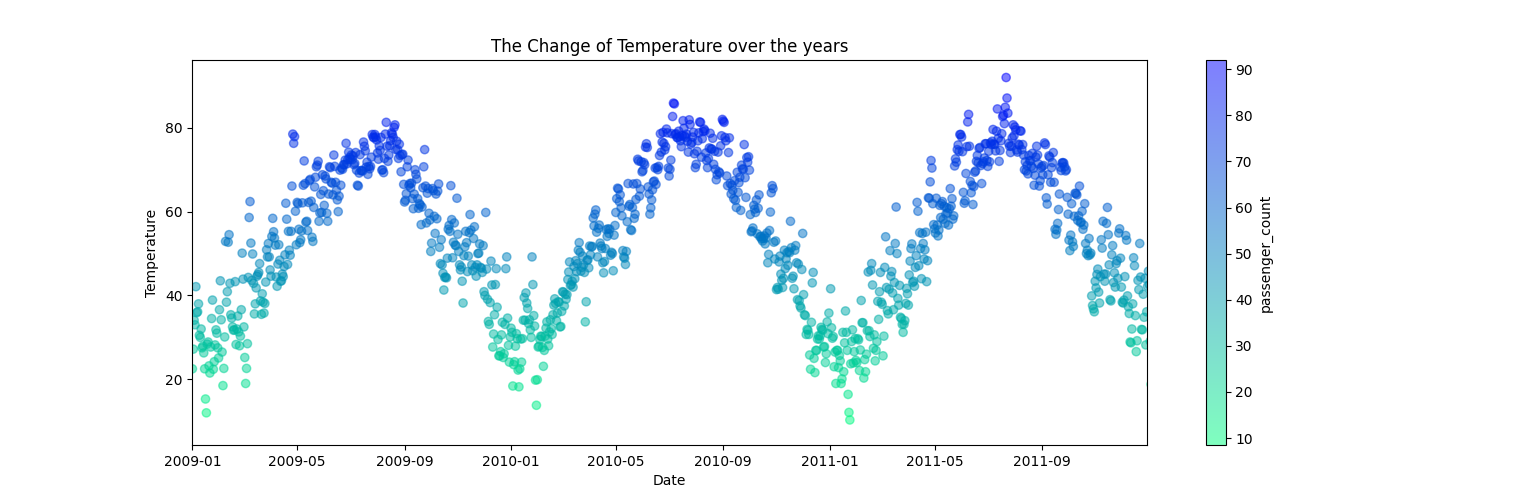
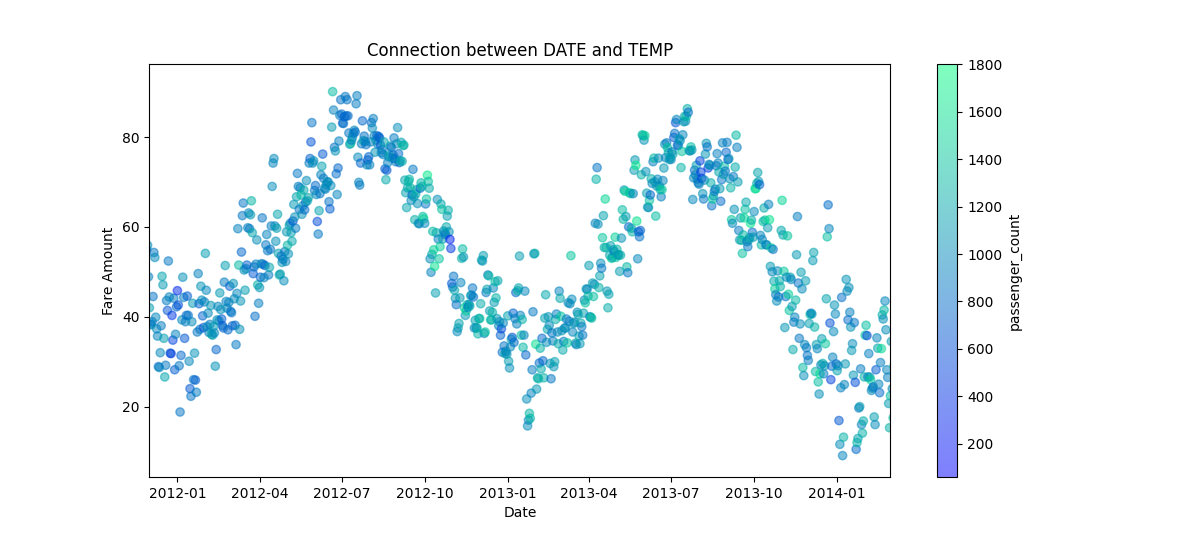
This is one of the most important steps when analyzing the data so we made many different manipulations on the data to fit our different visualizations and so we can get the desired results that we asked in our goal and objectives:

* When does the profit increase and how to make it higher.
* What If the reason behind delay in orders is the weather.
* What are the Uber passengers' number and its relation with the weather condition.

These objectives and many more were applied on different visualization and then we get the upcoming results and others.

* Increasing the profit throughout the whole year.
* Solutions for cars to perform better in any weather condition.
* Ways to keep the customer interested in Uber services.

**Some visualizations are listed below:**



And all the details along with documentation are labeled in the ZIP file as Visualization code.py

And the visualization are presented in the link from the file VisualPresentation.docx

**Conclusion:**

**The weather has a big impact on the performance of the Uber Drivers and the increase or** **decrease of the profit throughout the years.**

**References:**

<https://data.gov/>

<https://www.noaa.gov/>

<https://www.kaggle.com/>