Zain Goheer - 800936558

Premkumar Sreedhar Vemula - 801101045

Project 2 Report

We cleaned the data in the following order to ensure one section of the cleaning wouldn’t interfere with another:

1. Replace city data (lat, lon, inme, city, prov) - this geographical data does not relate to the numerical data so it is cleaned first and not referenced in the rest of the code. This was done by finding the mode of each column for each station which was placed into a separate dataframe, and all values in the main dataframe were overwritten with the mode to clean any anomalies. We found two extra anomalous themes outside of the required lat and lon cleaning:

* 31 entries for Station ID 178 are listed with an inme as A353 for between 1/1/2010 and 1/12/2010. The correct inme for this station is A333.
* 97 entries for Station ID 304 are listed with a city Santa Monica between 5/12/2012 and 5/16/12. The correct city name is Santa Teresa.

These anomalies were found using the .value\_counts() methods on each of the fields while analyzing the data before cleaning.

1. Here the feature handling is being done ie. fixing the 0 values, missing values and null values.

*Feature: Wind speed (wdsp)*

Handled wind speed based on wind gust and wind direction as well, ie. if either of the feature is available we have taken mean of wind speed based on year and imputed in wdsp.

For the large chunks of data where there are 0 values we have taken the mean for the whole column excluding the nulls and imputed the values in the column.

*Feature: Wind direction(wdct)*

Handled wind direction by imputing mean in all the 0 values as when there is wind speed there is a wind direction.

*Feature: Wind Gust(gust)*

Handled the wind gust, ie. wherever there was gust value made sure that the wind speed is available, wherever there is 0 filled it with mean of column excluding the null and 0 valued rows and left the null rows as is because it indicates there was no wind gust during that particular log.

*Feature: Air Pressure(stp)*

For stp feature's cleaning we impute the mean in all the rows that are 0s, null or outliers. The outliers found upon analysing the data were anything below 877.3 and above 1030.3. The mean is calculated per station as the stp varies based on the Altitude and replaced respectively in the column values. 0 is replaced in the similar fashion because the stp cannot be 0 on earth’s atmosphere.

*Feature: Air temperature(temp)*

For air temperature we have handled it with respect to the stations, ie. we took the mean of the values excluding the nulls and 0 for every station per year per month and replaced the means respectively in the 0 and nulls. For the data that has no values at all (this might be a reason because of the sensor error) in the dataset for a whole month we have computed the complete mean and imputed it.

*Feature: Dew PointTemperature (dewp)*

For dewp we have handled it similar to the Air Temperature, we left the -ve values as the temperature can be -ve.

*Feature: Relative Humidity(hmdy)*

The hmdy cannot be 0 as it is impossible for humidity to be 0 on earth surface, Water vapor is always present in the air, even if only in minute quantities. Hence we took a mean an impute in the column wherever there is 0 humidity.

1. Fixed precipitation anomalies - In this section we addressed 0’s and anomalies in the precipitation column. We then calculated the average between the last and next entry and used that value to fill the 0 or anomaly. If there was an anomalous amount of rainfall (according to wikipedia extreme rainfall is > 50mm in an hour, so we set the limit to 60) this was also overwritten with an average. We decided to find anomalies based on rainfall amount instead of hour-to-hour because severe rainfall can come and go within an hour. At the end of this section we cleared any remaining 0’s in prcp with null values and finalized this column.
2. The project instructions said temperature change of 20 degrees within an hour so we used a change of >= 20 degrees to determine an anomalous entry. We found the absolute value of the temperature difference between any consecutive entries, and if the change was greater than 20 degrees it was overwritten with the average of the previous and next recorded temperature.

Questions in order of rubric:

1. Precipitation was described in part 3 above
2. Zeroes in precipitation were addressed in part 3 above. Lone anomalous zeroes and big blocks of zeroes were addressed in part 2 above.
3. Addressed in Part 2
4. Addressed in part 4
5. Addressed in part 1
6. Addressed in part 1 and 2
   1. 31 entries for Station ID 178 are listed with an inme as A353 for between 1/1/2010 and 1/12/2010. The correct inme for this station is A333.
   2. 97 entries for Station ID 304 are listed with a city Santa Monica between 5/12/2012 and 5/16/12. The correct city name is Santa Teresa.
   3. Air pressure cannot be in 0 and there were around 241 records with 9999, and 28644 records with 0 and 1 record with outliers 1700, 1 record with outlier 2000 and 2 records with outlier 1800. Hence while handling these records we excluded the outliers and calculated the mean for every station and replaced respectively.