MidSem Project Evaluation

CSE 558 : Data Science

Group Name: Grade-ient Descent

Problem Statement

The aim of this project is to develop predictive models for two critical air pollutants, PM2.5 and PM10, using a comprehensive dataset encompassing hourly air quality measurements and meteorological data from 12 nationally-controlled air-quality monitoring sites in Beijing. The goal is to build a regression model for PM2.5 prediction and a classification model to assess the level of air quality based on PM10 concentrations. The project seeks to address the challenges of pollution monitoring and forecasting, contributing to improved public health, urban planning, and environmental policy-making.

Progress so far & Future plan of action

We performed an exploratory data analysis, and on the basis of that we set boundaries for our seasonal predictions and further analysis in the future. It also helps us study patterns in the data and set more informed hypotheses.

Further improve our seasonal analysis and use the improvements to aid the development for both models, regression on PM2.5 values and classification on PM 10, whose appropriate boundary values can also be determined by our current and future progress.

Challenges with raw data and our approach

- Due to the availability of our chosen dataset on kaggle we were initially apprehensive about choosing the given dataset as several ML models had been prepared already. But on further searching and observation, there weren’t any hypotheses regarding seasonal changes and their effects.
- Due to the data being a time series, missing values were abundant in the dataset. There were chunks of data values in the dataset missing. These were solved by substituting them with mean values.

Observing correlation between pollution levels and pollutants

We used Pearson’s method to find correlation coefficients

Target Variables	SO2	NO2	CO	O3
PM2.5 (sampled randomly)	0.483	0.658	0.762	-0.150
PM2.5 (entire data)	0.478	0.658	0.767	-0.151
PM10 (sampled randomly)	0.461	0.650	0.682	-0.112
PM10 (entire data)	0.459	0.645	0.684	-0.113

HYPOTHESIS TESTING

h0(null hypothesis): more than 21% of the days have unhealthy air quality(PM2.5 >= 150)

h1(alternate hypothesis): less than 21% of the days have unhealthy air quality(PM2.5 >= 150)

level of significance: 0.05

Using the Z-test for proportionality we carried out our hypothesis test.

Threshold for unhealthy Air Quality is AQi>=150

How was the number of 21% obtained: Chinese news websites have reported over 30+ experiencing good air quality 79% of the time("Pollution in China," n.d.). Our null hypothesis provides a conservative estimate.

Results

For random same of 10% size

The number of days with unhealthy air quality are: 6427

The total number of days are: 42077

The proportion of days with unhealthy air quality is: 0.1527437792618295

The z-score is: -28.835127064705695

The critical z-value is: -1.6448536269514729

Thus,we reject the null hypothesis

For Validation dataset of full size

The number of days with unhealthy air quality are: 63412

The total number of days are: 420768

The proportion of days with unhealthy air quality is: 0.15070537683474028

The z-score is: -94.43075721099315

The critical z-value is: -1.6448536269514729

Here too,We reject the null hypothesis confirming our sample result

HYPOTHESIS TESTING

H0 (Null Hypothesis): The pollution levels are independent of wind direction

H1 (Alternate Hypothesis): The pollutions levels are not independent of wind direction

Made the observed and expected table for the features above and then calculated the chi_squared value, degrees of freedom and p-value for the data.

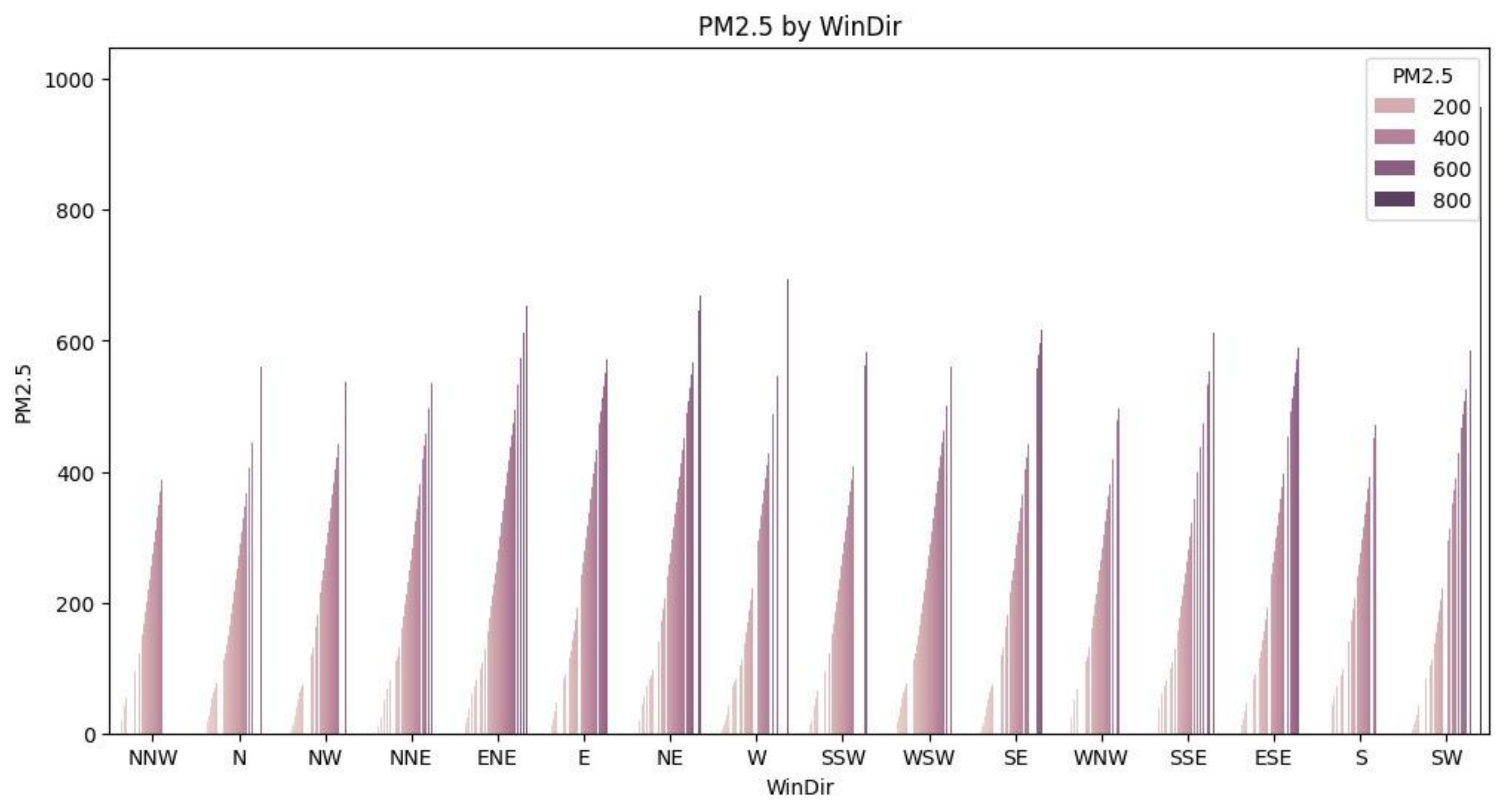
Initially, the test was done on 10% of the entire data sampled randomly; the results obtained were:

```
Chi-squared statistic: 642.798024316166; Degrees of freedom: 15; P-value: 2.82282638031e-122
```

The test was then performed on the entire dataset and the results obtained were as follows:

```
Chi-squared statistic: 6281.066330916687; Degrees of freedom: 15; P-value: 0.0
```

Thus, we reject the null hypothesis and conclude that the pollution levels depend on the direction of the wind.


PM2.5 by WinDir
PM2.5
1000
800
600
400
200
0
NNW
N
NW
NNE
ENE
E
NE
W
SSW
WSW
SE
WNW
SSE
ESE
S
SW
WinDir
PM2.5
200
400
600
800

HYPOTHESIS TESTING

h0(null hypothesis): The mean pollution levels are same on weekdays and weekends

h1(alternate hypothesis): The mean pollution levels are different on weekdays and weekends

Data preprocessing: Days were converted to either of 2 classifications either weekend or weekdays after which the data was grouped into the relevant arrays.

Test used:T-test for 2 means was used (assumption being that the standard deviation of both samples would be equal)

Differences in average values in sample dataset:

```
Weekday 78.776839 Weekend 84.010460 Name: PM2.5

Weekday 103.028795 Weekend 110.284191 Name: PM10
```

Results

For random same of 10% size

```
t-statistic for PM2.5: -6.0579843348265525
p-val for PM2.5: 1.3900455488946596e-09

t-statistic for PM10: -7.374132660327755 the
p-val for PM10: 1.684557461341611e-13
```

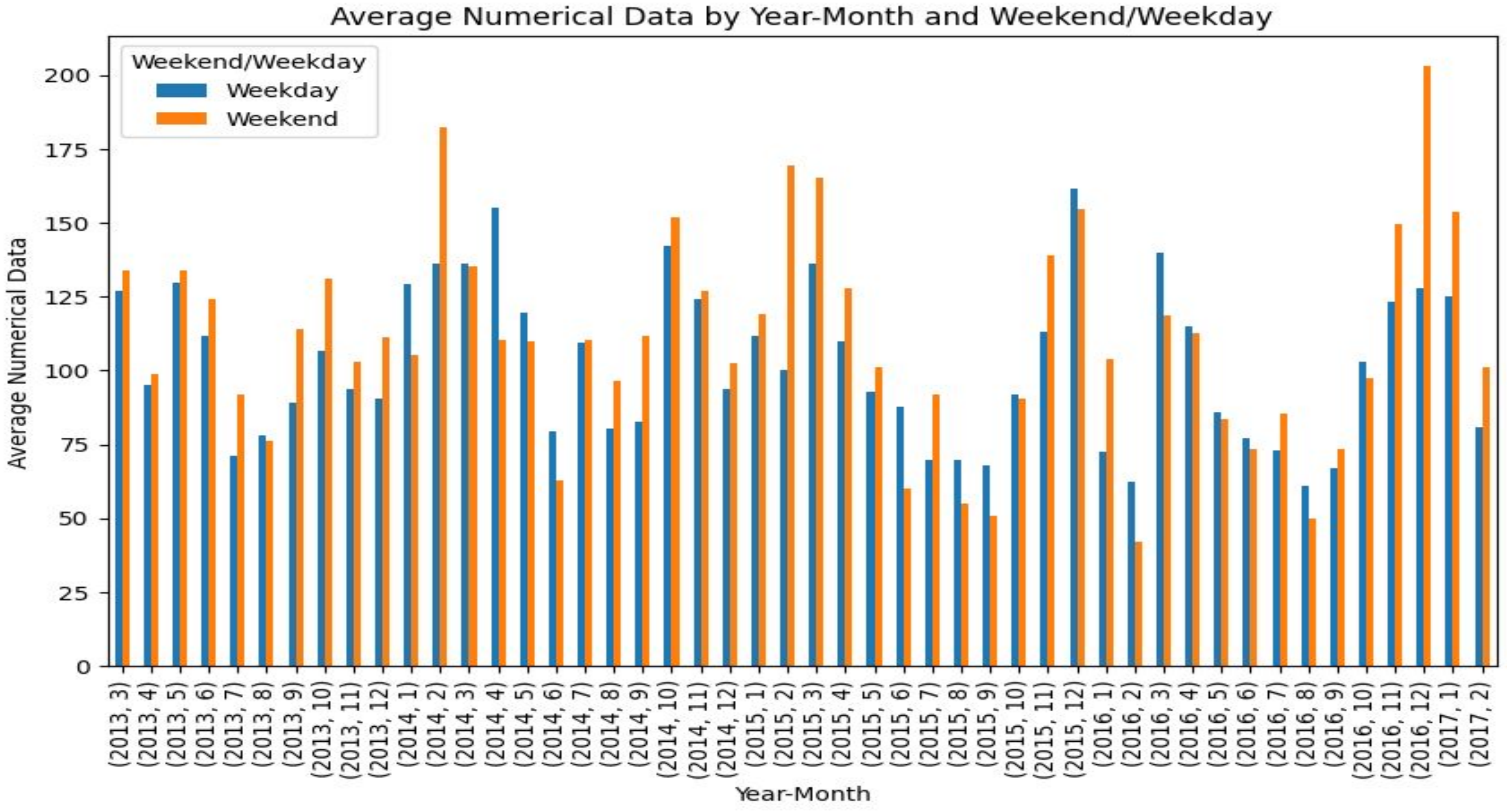
For p_values<0.05 we can argue that there is suitable evidence for a statistically difference between samples

For Validation dataset of full size

```
t-statistic for PM2.5: -21.543030823482038
p-val for PM2.5: 6.997977098029 23e-103

t-statistic for PM10: -25.810958164220636 the
p-val for PM10: 8.70030402394265e-147
```

This confirms our null hypothesis as the same p-values are being observed in both validation and test data


Average Numerical Data by Year-Month and Weekend/Weekday
Weekend/Weekday
Weekday
Weekend
Average Numerical Data
0
25
50
75
100
125
150
175
200
Year-Month
(2013, 3)
(2013, 4)
(2013, 5)
(2013, 6)
(2013, 7)
(2013, 8)
(2013, 9)
(2013, 10)
(2013, 11)
(2013, 12)
(2014, 1)
(2014, 2)
(2014, 3)
(2014, 4)
(2014, 5)
(2014, 6)
(2014, 7)
(2014, 8)
(2014, 9)
(2014, 10)
(2014, 11)
(2014, 12)
(2015, 1)
(2015, 2)
(2015, 3)
(2015, 4)
(2015, 5)
(2015, 6)
(2015, 7)
(2015, 8)
(2015, 9)
(2015, 10)
(2015, 11)
(2015, 12)
(2016, 1)
(2016, 2)
(2016, 3)
(2016, 4)
(2016, 5)
(2016, 6)
(2016, 7)
(2016, 8)
(2016, 9)
(2016, 10)
(2016, 11)
(2016, 12)
(2017, 1)
(2017, 2)

HYPOTHESIS TESTING

H0: $\mu = \mu_0$;

H1: $\mu \neq \mu_0$;

Test used:Z-test for population means was used with 0.05 level of significance.

```
Computing Z values and Sample means :
1. PM 2.5 Sample Mean = 80.16539648084368
PM 2.5 Population Mean = 79.76656545075912 and Population Std Deviation =
80.00980321950516
Z Score is : 0.7230160918300503


2. PM 10 Sample Mean = 105.23972918001382
PM 10 Population Mean = 104.61350045300435 and Population Std Deviation =
91.09328567144532
Z Score is : 0.9971232537182779
```

HYPOTHESIS TESTING

H0: $\sigma^2 = \sigma^2_0$;

H1: $\sigma^2_0 \neq \sigma^2_0$;

Test used: χ^2-test for population means was used with 0.05 level of significance.

```
Computing Chi-Square values and Sample Std Deviation :
1. PM 2.5 Sample Variance = 6386.154769573281
PM 2.5 Population Variance = 6401.568611223938
Chi Score is : 20986.34663572357

2. PM 10 Sample Variance = 8358.115668764962
PM 10 Population Variance = 8297.986694419544
Chi Score is : 21189.438570931336
Since the degree of freedom is 21037 thus critical value is
21375.524713112194
```

HYPOTHESIS TESTING

H0: $r = r_1$;

H1: $r \neq r_1$;

Test used: T-test for correlation coefficient was used with 0.05 level of significance.

Feature / Target	PM 2.5	PM 10

Rainfall	r = -0.022629301286238798
	t = -2.692564478919192e-06
		r = -0.030312135436531395
		t = -3.1319216717855362e-06

Pressure	r = 0.028057336322297562
	t = 2.3249102374396277e-07
		r = -0.008932531626347839
		t = -6.427375989723267e-08

Temperature	r = -0.14147442466355364
	t = -1.0658730791667352e-06
		r = -0.10343233592027638
		t = -6.766799079527533e-07

HYPOTHESIS TESTING

H0 (Null Hypothesis): The PM10 particles in the air during different seasons are independent to each other.

H1 (Alternate Hypothesis): We cannot separate the analysis of the data based on seasonality.

Test used: Chi squared-test for independence of the attributes with 0.5 level of significance.

MOTIVE:

The main objective of the project is to perform seasonal analysis on the time data. Do to so, we need to check that there is no relation between the trends of the residuals in air in different seasons.

P - VALUE	SPRING	SUMMER	AUTUMN	WINTER
SPRING	0.00000000e+00	1.09873056e-82	1.05002779e-17	1.38752525e-09
SUMMER	1.09873056e-82	0.00000000e+00	3.46554451e-29	3.73745217e-23
AUTUMN	1.05002779e-17	3.46554451e-29	0.00000000e+00	4.07855754e-43
WINTER	1.38752525e-09	3.73745217e-23	4.07855754e-43	0.00000000e+00

RESULT: In any of the cases, there is no significant evidence to reject the null hypothesis. Hence, the pollution trend of each season is independent of the other.

Thank you