Assignment 1

Data Exploration and Regression Analysis

Semester 1 2023

Student Name: Samantha Heuss
Student ID: 21141141
PAPER NAME: Foundations of Data Science

PAPER CODE: COMP615

Due Date: Friday 14 April 2023 (midnight)

TOTAL MARKS: 100

INSTRUCTIONS:

1. **The following actions** may be deemed to constitute a breach of the General Academic **Regulations Part 7: Academic Discipline**,
- Communicating with or collaborating with another person regarding the Assignment
- Copying from any other student work for your Assignment
- Copying from any third-party websites unless it is an open book Assignment
- Uses any other unfair means

2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately**
3. **Attach your code for all the datasets in the appendix section.**

Table of Contents

The Effect of Wind Patterns on the Concentration of $PM_{2.5}$

1 Introduction

This report will focus on the relationship between wind patterns and the changes in levels of $PM_{2.5}$ present in the Beijing atmosphere. $PM_{2.5}$ are tiny particles that pollute the air and reduce the visibility, causing the air to become hazy. Additionally, higher levels pose a risk to people's health. [1] The aim of this report is to establish, firstly, if this relationship does indeed exist. Then secondly, determine the nature of the relationship and what this means for the future of Beijing as weather patterns change due to climate change and rising global temperatures. These areas of focus can be condensed into a select few questions to direct the research towards a clear conclusion: "Is there a relationship between wind patterns and the concentration of the $PM_{2.5}$? How do the changes in one of these variables affect the other? What are the possible implications of these findings for the future of air quality management in Beijing?". The Beijing $PM_{2.5}$ Data dataset will be used to explore this correlation. From this dataset, the main attributes focused on are the PM2.5 concentration (pm2.5) measured in ug/m^3, Combined wind direction (cbwd), and Cumulated wind speed (Iws) measured in m/s.

There are a few assumptions the dataset must meet for the conclusions made to be considered reliable. Firstly, a sufficient sample size must be used. As the data is a set of records obtained hourly over the course of five years, the sample size can be regarded as suitable. Additionally, the dataset needs to be obtained from a reliable source. The data has been collected and recorded by the Guanghua School of Management, which is a Center for Statistical Science at Peking University [2], and therefore, a reliable source of information to use in this report. It is equally important to address the quality of data. While it is assumed that each hourly record of data is independent of the others as to not affect or influence the record of another, a missing value for example could affect this. Furthermore, it is assumed there is no high correlation between two or more independent variables. Doing so will cause multicollinearity to occur, thereby making it hard to distinguish between the effects of each independent variable on the dependent variable. Additionally, the assumption of homoscedasticity is achieved when the variability (the spread) of the dependent variable is the same across all levels of the independent variable.

2 Data Exploration

The purpose of this dataset was to capture $PM_{2.5}$ concentration and meteorological data in Beijing. This dataset features data collected hourly by the Guanghua School of Management, part of the Center for Statistical Science at Peking University, between January 1st, 2010, and December 31st, 2014. [2] The data was recorded in two separate locations; the $PM_{2.5}$ data was captured at the US Embassy in Beijing, whilst all meteorological data was captured at the Beijing capital International Airport. [2] In total, there are 13 attributes in the dataset: row number, year of data in this row, month of data in this row, day of data in this row, hour of data in this row, $PM_{2.5}$ concentration (ug/m^3), dew point (degrees Celsius), temperature (degrees Celsius), pressure (hPa), combined wind direction, cumulated wind speed (m/s), cumulated hours of snow (hours), and cumulated hours of rain (hours). There are exactly 43,824 instances in the dataset.

Importing this dataset into python, the describe function produces a table of all the attributes, and their continuous numerical features (see table of continuous numerical features). Regarding my research

focus, the main columns in the dataset I will be focusing on are pm2.5 ($PM_{2.5}$ concentration), cbwd (combined wind direction), and Iws (cumulated wind speed). Observing the count row reveals the missing data in the dataset to only be in the $PM_{2.5}$ concentration column, as the count differs from the count of all other attributes in the dataset. These missing values are represented as "NA". The cbwd (combined wind direction) column is not included within the describe table, as the data in this column is non-numerical data. The data is instead represented as the direction of the wind such as NW, SE, etc. as a text datatype; the value cv represents a calm wind with speed so little it cannot be recorded as a direction. The pm2.5 column has a mean of 98.6 (3 sig. figs), a standard deviation of 92.1 (3 sig. figs), a minimum of 0.00, and maximum of 994.0. For context, the World Health Organisation recommends a concentration of $PM_{2.5}$ to be no greater than 15 $\mu g/m^3$ for a period of 24 hours. [3] The Iws column has a mean of 23.9 (3 sig. figs), a standard deviation of 50.0 (3 sig. figs), a minimum of 0.45, and a maximum of 586 (3 sig. figs). All other values can be observed in the describe table (see table of continuous numerical features).

2.1 Data Preparation

There are many ways to address the missing data in the pm2.5 column. One approach is, for the pm2.5 column, the available data can be modelled, and the missing values an estimate based on the trend(s) observed in these graph(s). However, due to the vast number of instances in the dataset, and similarly the large quantity of missing values to resolve, the better option here is to remove the rows in the dataset that contain missing data. Though the outcome will cause small gaps to arise in the consistency of the data, the far greater number of records that do contain data for all attributes at that time will compensate for this and enable a fair conclusion.

2.2 Data Analysis

To further explore the dataset and understand its trends, the data can be visualised into various graphs. The graph titled "Cumulated Wind Speed vs PM2.5 Concentration" (see Fig. 1) is a line graph featuring the titled attributes. When the concentration is 0, the cumulated wind speed nears 200 m/s. From there, the cumulated wind speed rapidly drops until it reaches the graph's asymptote, keeping a constant trend of the cumulated wind speed around 15 whilst the $PM_{2.5}$ concentration is between approximately 50 and 1000. This suggests that a strong wind speed can reduce the concentration of $PM_{2.5}$ until a certain point. The "PM2.5 Concentration vs Combined Wind Direction" graph (see Fig. 2) and the "Cumulated Wind Speed vs Combined Wind Direction" graph (see Fig. 3) can be observed together. The NW wind has the highest speed, and the lowest $PM_{2.5}$ concentration, whilst the cv wind has the lowest speed and highest $PM_{2.5}$ concentration. The SE and NE values differ from this observed trend. Beijing's north-east location in China is likely to cause this change, as the easterly winds push the particles towards Beijing, thereby increasing the $PM_{2.5}$ concentration. [3] "PM2.5 Concentration for each Month" (see Fig. 4) and "Cumulated Wind Speed for each Month" (see Fig. 5) are two line graphs that visualise the cumulated wind speed and $PM_{2.5}$ concentration for each month. The PM2.5 concertation shows an increase around the months of February, June, July, and October. Contrarily, the wind speed decreases during the months of February, and June through to October.

Before visualising the dataset, the data had been cleaned by applying complete case analysis as explained above. One assumption outlined in the introduction was a sufficient sample size was used. After the removal of this data, the number of instances is reduced to 41,756, so the assumption can still be made. Additionally, there is no evidence of violation of the assumptions of homoscedasticity or presence of multicollinearity. The data therefore remains to be reliable.

3 Correlation Analysis

Observing the whole dataset, independent variables dew point and temperature both have a strong negative correlation with pressure, -0.78 and –0.83 respectively, indicating the presence of multicollinearity. Having this makes it difficult to interpret each of these individual independent variable's effects on the changing of pressure. To resolve, one solution is to remove one of the independent variables from the dataset. As the focus of the report is how the changes in wind speed and direction affect the $PM_{2.5}$ concentration, all three of these variables could be removed from the dataset.

Using only the variables deemed relevant to the research focus, $PM_{2.5}$ and cumulative wind speed have a weak negative correlation of –0.25 (see Fig. 8). Based on this analysis only, there is no sign of multicollinearity between these independent variables. Depending on the context, the $PM_{2.5}$ concentration can also be considered a dependent variable. It can then be said there is also no multicollinearity present between these independent and dependent variables. To further establish if there is a presence of multicollinearity, additional steps must be taken. To do so, a scatter matrix has been plotted (see Fig. 9). The graph of $PM_{2.5}$ concentration against cumulated wind speed and vice versa in the scatter matrix, both have a negative association. Additionally, the relations in both these scatter graphs are non-linear, rather the points are scattered around a curve. These therefore support the claims from the heatmap that there is no multicollinearity, as variables said to have a non-linear correlation will have a very low correlation. [4] The multicollinearity assumption remains unviolated.

Expanding this investigation to include additional attributes allows a wider picture to be drawn, and therefore a better understanding of what is happening and why. Air pressure has a direct effect on wind patterns. A pressure gradient is created when one part of the atmosphere is significantly lower than the areas surrounding it. Wind occurs as air moves into the lower pressure part of the pressure gradient to resolve it. The greater the pressure gradient, the stronger the wind. [5] Observing the heatmap (see Fig. 6), pressure and cumulated wind speed too has a weak correlation of 0.19. Unlike the attributes mentioned above, this is a positive association. The shape is not linear, rather the scatter points are clumped together around where x is 0 and spreads as x increases. This graph therefore has heteroskedasticity and violates the assumption of homoscedasticity. This occurs between two independent variables though, the scatterplot between pressure and $PM_{2.5}$ concentration doesn't have this issue. Though this graph is too non-linear, there is no evidence of heteroskedasticity between these independent and dependent variables, so there is no reason to resolve this issue at this time. Subsequently, no change has been made to the sample size or origin of data, so these assumptions remain valid.

4 Linear Regression

Five ols (ordinary least squares) models have been produced to explore the linear regression in the dataset, using the attributes deemed relevant for the focus of the research (see Figs. 11 through 16). Models 1, 2, 3, and 4 all use $PM_{2.5}$ concentration as the dependent variable. In model 1, the independent variable is the cumulated wind speed, in model 2 it is pressure, and models 3 and 4 are the cumulation of these two. Model 5 uses pressure as the independent variable to determine its effect on the dependent variable cumulated wind speed.

Table 1. Summary table of r-squared values and p-values.

	R-squared	Adjusted R-squared	P-value
Model 1	0.061	0.061	0.000

Model 2	0.02	0.02	0.000
Model 3	0.061	0.061	0.002
Model 4	0.061	0.061	0.002
Model 5	0.034	0.034	0.000

5 Statistical Inference

In the linear regression, the r-squared and adjusted r-squared value for models 1, 3, and 4 were all 0.061. The adjusted r-squared value will only differ from the r-square value when more dependent values are included into the ols (ordinary least squares) model. The value for these models is very low, which implies the models do not reflect the larger proportion of variation in the dependent variable, $PM_{2.5}$ concentration. Taking into account the results from correlation analysis obtained prior to this, there is no multicollinearity in these models. This can be further proved using the r-squared value to calculate the Variance Inflation Factor (VIF) which determines the presence of multicollinearity. The VIF value where the r-squared value is 1.065 (3 sig. figs). This value is close to 1, and therefore indicates no multicollinearity. [6]

$$1 / (1 - r^2) = \text{Variance Inflation Factor} \quad (1)$$

The r-squared and adjusted r-squared value for model 2 is 0.002. Similarly, this value is also low, and does not reflect the larger proportion of variation in the dependent variable. The calculated VIF value for this model is 1.00 (3 sig. figs), and therefore too has no indication of multicollinearity. In the correlation analysis, there was no indication of multicollinearity here, and this therefore supports this result. As established in the correlation analysis, the only presence of multicollinearity in the dataset occurs between the independent variables pressure and dew point against the dependent variable $PM_{2.5}$ concentration. These values are not used in this research, and do not affect the outcome. However, the r-squared/adjusted r-squared values significantly low values indicate other values in the dataset may have a far greater effect on the $PM_{2.5}$ concentration than wind, and subsequently pressure's effect on wind then $PM_{2.5}$ concentration.

The confidence levels of 0.05 are not exceeded by any of the models, no there is no statistically significant values between any of the attributes used in these models. The assumptions of sample size and reliable sources continue to remain unviolated as no changes have been made to either of these. The homoscedasticity assumption is not affected here, but if the factors of research were to change as suggested, this too could become violated.

6 Conclusion

The question proposed in the introduction to direct the research to a clear conclusion were "Is there a relationship between wind patterns and the concentration of the $PM_{2.5}$? How do the changes in one of these variables affect the other? What are the possible implications of these findings for the future of air quality management in Beijing?" Based on the dataset, and the models created above, these questions can now be answered.

The graphs produced initially whilst visualising the data indicate that wind does influence the concentration of $PM_{2.5}$. The two line graphs (see Fig. 4 and Fig. 5) are a clear indication of this; during the months the wind is increasing, the $PM_{2.5}$ concentration is decreasing, and vice versa. This is contradicted when carrying out a correlation and linear regression analysis, as these indicate that there is no high correlation between the dependent and independent variables, nor is the r-squared values high enough to reflect the larger

proportion of variance in the dependent variables. As these forms of analysis are more reliable, what appears to be wind speeds effecting the $PM_{2.5}$ concentrations may be purely coincidental. Rather, other attributes may be affecting this concentration. It is my recommendation then that these other attributes be further investigated. Temperature and dew point for example have a strong correlation with pressure, which could be one focus moving forward. Another could be a deeper investigation into all meteorological factors in association with seasonal changes and what this effect has on $PM_{2.5}$ concentrations, not solely focusing on wind patterns alone.

7 Appendix

For all graphs, tables, and code, please refer to the additional markdown file provided.

Table 2. Summary of all equations, figures, and tables.

Equation 1	Equation of Variance Inflation Factor
Figure 1	Line graph of Cumulated Wind Speed vs PM2.5 Concentration
Figure 2	Bar graph of PM2.5 Concentration vs Combined Wind Direction
Figure 3	Bar graph of Cumulated Wind Speed vs Combined Wind Direction
Figure 4	Line graph of PM2.5 Concentration for each Month
Figure 5	Line graph of Cumulated Wind Speed for each Month
Figure 6	Heatmap of dataset
Figure 7	Scatter plot matrix of dataset
Figure 8	Heatmap of condensed dataset
Figure 9	Scatter plot matrix of condensed dataset
Figure 10	Scatter graph of Pressure vs Cumulated Wind Speed
Figure 11	Linear regression model of Cumulated Wind Speed and $PM_{2.5}$ concentration
Figure 13	Linear regression model of Pressure and $PM_{2.5}$ concentration
Figure 14	Linear regression model of Cumulated Wind Speed, Pressure and $PM_{2.5}$ concentration
Figure 15	Linear regression model of Cumulated Wind Speed, Pressure and $PM_{2.5}$ concentration
Figure 16	Linear regression model of Cumulated Wind Speed and Pressure
Table 1	Summary table of r-squared values and p-values.
Table 2	Appendix summary table of all equations, figures, and tables

8 Bibliography

1. Adhikari, A., DeNero, J., & Wagner, D. (2021, April 30). *Computational and inferential thinking*. 15.1. Correlation - Computational and Inferential Thinking. Retrieved April 13, 2023, from https://inferentialthinking.com/chapters/15/1/Correlation.html

2. *Department of Health*. Fine Particles (PM 2.5) Questions and Answers. (2018, February). Retrieved April 11, 2023, from https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#:~:text=Fine%20particulate%20matter%20(PM2.5,hazy%20when%20levels%20are%20elevated
3. Gellert, A. (2019, March 2). *How does pressure affect wind?* Sciencing. Retrieved April 13, 2023, from https://sciencing.com/pressure-affect-wind-23262.html
4. Stats NZ. (2022, October 27). *$PM_{2.5}$ concentrations: Stats NZ*. $PM_{2.5}$ concentrations | Stats NZ. Retrieved April 5, 2023, from https://www.stats.govt.nz/indicators/pm2-5-concentrations/#:~:text=Currently%20no%20standard%20is%20set,World%20Health%20Organization%2C%202021).
5. Stephanie. (2020, December 16). *Variance inflation factor*. Statistics How To. Retrieved April 16, 2023, from https://www.statisticshowto.com/variance-inflation-factor/
6. UCI Machine Learning Repository: Beijing PM2.5 Data Data Set. (n.d.). Retrieved March 23, 2023, from https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data