**VPD Report**

Introduction ………………………………………………………….4

6 Suicides Cases 4

Katie Meyer 5

Robert Martin…………………………………………………….6

Jayden Hill…………………………………………………………..7

Sarah Schulze 8

Caitlin Kelly 9

Lauren Bernett………………………………………………….10

What is VPD………………………………………………………….11

VPD Analysis for all six cities 12

VPD Actual Dataset ……….……………………………………..12

VPD Metadata ………………………………………………........12

VPD Analysis for Palo Alto 13

Paulo Alto Dataset Head ………………………… 13

Palo Alto Dataset Description 13

Palo Alto Rate of Change in VPD per Year 14

Palo Alto Correlation Heat Map 16

Linear Regression for Palo Alto 17

Interpreting OLS Regression Results 18

Dependent Variable 18

Model 19

Number of observations 19

Degree of freedom(df) of residuals 19

Coefficient term 19

Standard error of parameters 19

t statistics 20

p values 21

Confidence intervals 21

R squared value 21

F statistic 22

VPD Analysis for Johnson City 23

Johnson City Dataset Head ……………………. 23

Johnson City Dataset Description 23

Johnson City Rate of Change in VPD per Year 24

Johnson City Correlation Heat Map 26

Linear Regression for Johnson City 27

Interpreting OLS Regression Results 28

Dependent Variable 28

Model 29

Number of observations 29

Degree of freedom(df) of residuals 29

Coefficient term 29

Standard error of parameters 29

t statistics 30

p values 31

Confidence intervals 31

R squared value 31

F statistic 32

VPD Analysis for Marquette 33

Marquette Dataset Head ……………………….. 33

Marquette Dataset Description 33

Marquette Rate of Change in VPD per Year 34

Marquette Correlation Heat Map 36

Linear Regression for Marquette 37

Interpreting OLS Regression Results 38

Dependent Variable 38

Model 39

Number of observations 39

Degree of freedom(df) of residuals 39

Coefficient term 39

Standard error of parameters 39

t statistics 40

p values 41

Confidence intervals 41

R squared value 41

F statistic 42

VPD Analysis for Madison 43

Madison Dataset Head ………………………….. 43

Madison Dataset Description 43

Madison Rate of Change in VPD per Year 44

Madison Correlation Heat Map 46

Linear Regression for Madison 47

Interpreting OLS Regression Results 48

Dependent Variable 48

Model 49

Number of observations 49

Degree of freedom(df) of residuals 49

Coefficient term 49

Standard error of parameters 49

t statistics 50

p values 51

Confidence intervals 51

R squared value 51

F statistic 52

VPD Analysis for Weyers Cave 53

Weyers Cave Dataset Head ……………………… 53

Weyers Cave Dataset Description 53

Weyers Cave Rate of Change in VPD per Year 54

Weyers Cave Correlation Heat Map 56

Linear Regression for Weyers Cave 57

Interpreting OLS Regression Results 58

Dependent Variable 58

Model 59

Number of observations 59

Degree of freedom(df) of residuals 59

Coefficient term 59

Standard error of parameters 59

t statistics 60

p values 61

Confidence intervals 61

R squared value 61

F statistic .62

VPD Analysis for Philadelphia………………………..........63

Philadelphia Dataset Head ………………………… 63

Philadelphia Dataset Description 63

Philadelphia Rate of Change in VPD per Year 64

Philadelphia Correlation Heat Map 66

Linear Regression for Philadelphia 67

Interpreting OLS Regression Results 68

Dependent Variable 68

Model 69

Number of observations 69

Degree of freedom(df) of residuals 69

Coefficient term 69

Standard error of parameters 69

t statistics 70

p values 71

Confidence intervals 71

R squared value 71

F statistic 72

Code Explanation………………………………………………….73

Data Extraction………………………………………………….73

Data Description……………………………………………….75

Data Visualization …………………………………………….76

Data Analytics using Linear Regression Model….81

Results…………………………………………………………………87

What did I learn…………………………………………………..87

Was I successful…………………………………………………..87

Introduction

Environmental Conditions and Health an Analysis of VPD and Humidity in 6 Student Athlete Suicides.

Recent the Journal of the American Medical Society September 2022, published an urgent plea for action to combat the delirious effects of Global Warming and health. We started the process of studying VPD in the Winter of 2018. We were motivated by an outbreak of C. Auris which took the lives of four patients in a nursing home in New Jersey. Our approach was to find a method of measuring humidity that could explain the mysterious occurrences of C. Auris infections. We decided that the common index used in greenhouses could be a valuable index of patterns of humidity and temperature.

What we found was that VPD changed significantly during certain months. For example, in 2019 the VPD for October was significantly lower than in September, Average 1.441 vs 1.1212, t=5.09 p value <.001. We conducted several other t-tests on our data set and substantiated the findings that VPD fluctuates significantly on a month-to-month basis. Furthermore, given its ability to cause the propagation of vegetation we hypothesize that it also can encourage the growth of fungi and mold (see references).

In March and April (2019 to 2022) there were three student Athlete Suicides:

* Katie Meyer (March 11th, 2022)
* Robert Martin (April 1st, 2022)
* Jayden Hill (April 3rd, 2022)
* Sarah Shulze (April 13th, 2022)
* Kelly Caitlin (March 7th, 2019)
* Lauren Bernett (April 26th, 2022)

A close up of a person's eyes

Description automatically generated with medium confidence

**Katie Meyer** was born on January 20, 2000. She was an American Soccer player who played as a goalie for Stanford University’s Cardinal Women’s soccer team. She started at Stanford University in 2018, getting her degree in International Relations with a minor in History. The 22-year-old was found dead in her on-campus residence, and authorities determined she died by suicide. Mr. and Mrs. Meyer believed that a pending disciplinary hearing may have been a factor.

A close up of a chicken

Description automatically generated with medium confidence

Robert Martin was a graduate student at SUNY Binghampton New York. He played Lacrosse for his University as a Goalkeeper. He earned first team all leagues honor. He died on April 1st, 2022.



**Jayden Hill** age 19 died on April 3rd, 2022. She was born in Monrovia, Liberia. She was pursuing a degree in Political Science with a concentration in pre-law at Northern Michigan University. She was a Track Athlete for NMU Wildcat Athletics.

A close up of a person's eyes

Description automatically generated with medium confidence

**Sarah Schulze** committed suicide this spring was age 21 on April 13th, 2022. She ran indoor and outdoor tracks. She attended the University of Wisconsin in Madison. She earned Academic big ten honors in 2020 and 2021. She was an intern for the Wisconsin legislature and volunteered as a poll worker for the 2020 presidential elections. While attending Oak Park High School in California she won the Division 3 1600- and 3200-meter finals at El Camino College.

A close up of a person's face

Description automatically generated with medium confidence

**Catlin Kelly** died on March 7th,2019 by suicide at age 23. She was born in St. Paul Minnesota. She died in her dorm at Stanford University. She studied computational and mathematical engineering. She played the violin. She won 34 medals, 34 gold, 7 silver and 1 bronze in major international competitions. On January 5th, 2019, Kelly had a concussion due to a fall during training.

As a result of these mysterious suicides in the Spring of 2019 and 2022, we began investigating the VPD in the cities where these high-profile suicides occurred.

A close-up of a person smiling

Description automatically generated

**Lauren Bernett** is a Softball Player representing her University James Madison University in Wisconsin. She was born in Virginia, United States. She was a Standout player for the University and was awarded the CAA softball player of the year. She died of a heart attack on April 26th, 2022.

What is VPD?

VPD stands for Vapor Pressure Deficit, i.e., difference or deficit between the amount of moisture in the air and how much moisture the air can hold when it’s saturated

In simple words we use VPD to determine if the weather is too moist or too dry.

Keys Highlights of VPD:

VPD is calculated using two important factors viz. Temperature and Relative Humidity.

Temperature is Directly Proportional to VPD Score and Humidity is Inversely Proportional to VPD Score.

Formula:

(6.1078\*exp (Temp/(234.175 + Temp)\*17.08085)\*(1- Humidity/100))/10

VPD Score Range:

0.8 - 1.2 Kilo Pascal: Ideal Range

Below 0.8 Kilo Pascal: Too Moist

Above 1.2 Kilo Pascal: Too Dry

VPD Analysis for all Six Cities

* Palo Alto
* Johnson City
* Marquette
* Madison
* Weyers Cave
* Philadelphia

VPD Actual Dataset:

<https://wwo-bulk.s3.amazonaws.com/hwd_order_2487.zip>

VPD Metadata:

Table

Description automatically generated

VPD Analysis for Palo Alto

Paulo Alto Dataset Head:

Table

Description automatically generated

Palo Alto Dataset Description:

Table

Description automatically generated

Palo Alto Rate of Change in VPD per Year

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart

Description automatically generated

Palo Alto Correlation Heat Map

A picture containing graphical user interface

Description automatically generated

Linear Regression for Palo Alto

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results

Graphical user interface

Description automatically generated

Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0289 and **humidity** decreases by 0.0169. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.00042466**

**standard error(humidity) = 0.0004077**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0289 and se(b1) = 0.00042466, So,

t = (0.0289 – 0) / 0.00042466 = **68.054**

Similarly, b2= -0.0169, se(b2) = 0.0004077

                   t = (-0.0169 – 0) / 0.0004077 = **-41.45**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (68.054 and -41.45) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0289, s.e(b1)=0.00042466, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.028, 0.03). The C. I for B2 is approx. (-0.018, -0.016).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **99.3 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.993/2) / (0.007/130) = 9738.

You can calculate the probability of F >9738 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.536 and 3.361 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

VPD Analysis for Johnson City

Johnson City Dataset Head:

Table

Description automatically generated with medium confidence

Johnson City Dataset Description:

Table

Description automatically generated

Johnson City Rate of Change in VPD per Year

Chart, scatter chart

Description automatically generated

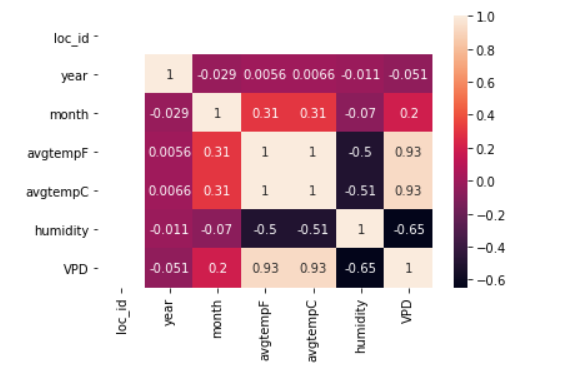
Table

Description automatically generated

Chart

Description automatically generated

Johnson City Correlation Heat Map

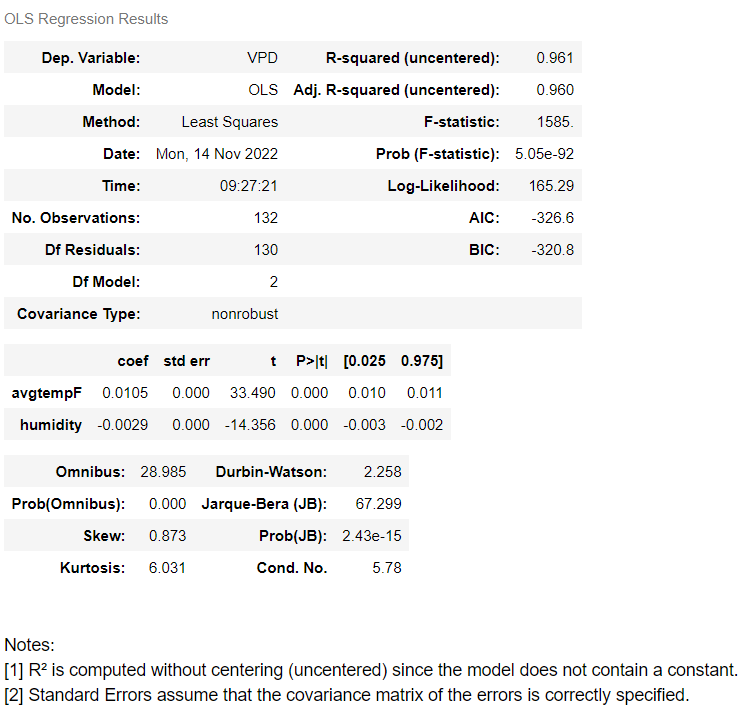


Linear Regression for Johnson City

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results



Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0105 and **humidity** decreases by 0.0029. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.0003135**

**standard error(humidity) = 0.000202**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0105 and se(b1) = 0.0003135, So,

t = (0.0105 – 0) / 0.0003135 = **33.49**

Similarly, b2= -0.0029, se(b2) = 0.000202

                   t = (-0.0029 – 0) / 0.000202 = **-14.356**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (33.49 and -14.356) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0105, s.e(b1)=0.0003135, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.010, 0.011). The C. I for B2 is approx. (-0.003, -0.002).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **96.1 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.961/2) / (0.039/130) = 1585.

You can calculate the probability of F >1585 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.873 and 6.031 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

VPD Analysis for Marquette

Marquette Dataset Head:

Table

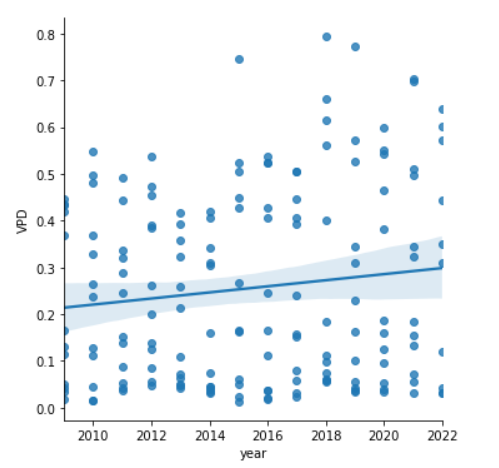
Description automatically generated

Marquette Dataset Description:

Table

Description automatically generated

Marquette Rate of Change in VPD per Year





Chart, line chart

Description automatically generated

Marquette Correlation Heat Map

A picture containing graphical user interface

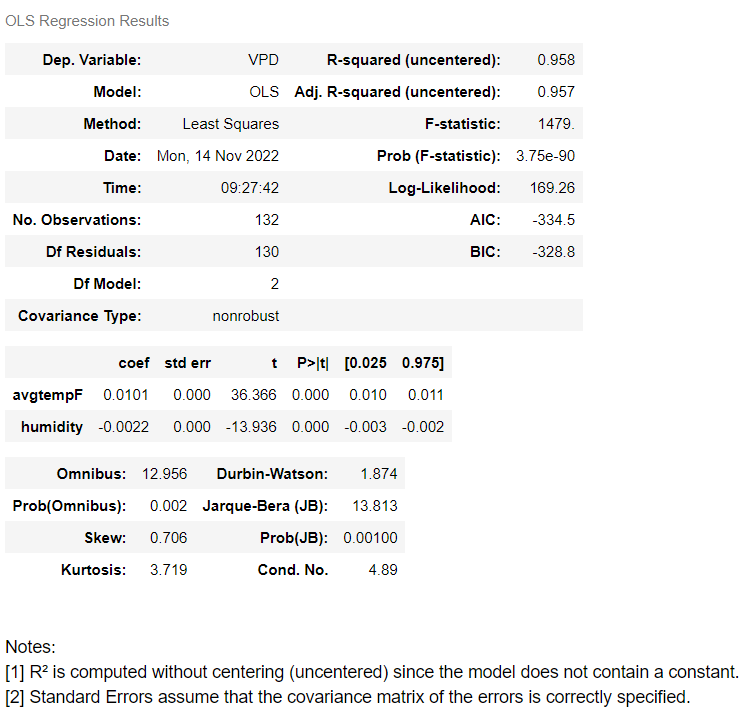
Description automatically generated

Linear Regression for Marquette

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results



Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0101 and **humidity** decreases by 0.0022. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.0002777**

**standard error(humidity) = 0.00015786**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0101 and se(b1) = 0.0002777, So,

t = (0.0101 – 0) / 0.0002777 = **36.366**

Similarly, b2= -0.0022, se(b2) = 0.00015786

                   t = (-0.0022 – 0) / 0.00015786 = **-13.936**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (36.366 and -13.936) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0101, s.e(b1)=0.0002777, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.010, 0.011). The C. I for B2 is approx. (-0.003, -0.002).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **95.8 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.958/2) / (0.042/130) = 1479.

You can calculate the probability of F >1479 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.536 and 3.361 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

VPD Analysis for Madison

Madison Dataset Head:

Table

Description automatically generated with medium confidence

Madison Dataset Description:

Table

Description automatically generated

Madison Rate of Change in VPD per Year

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart

Description automatically generated

Madison Correlation Heat Map

Graphical user interface

Description automatically generated with low confidence

Linear Regression for Madison

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results

Graphical user interface

Description automatically generated

Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0087 and **humidity** decreases by 0.0018. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.000321**

**standard error(humidity) = 0.0002015**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0087 and se(b1) = 0.000321, So,

t = (0.0087 – 0) / 0.000321 = **27.096**

Similarly, b2= -0.0018, se(b2) = 0.0002015

                   t = (-0.0018 – 0) / 0.0002015 = **-8.933**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (27.096 and -8.933) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0087, s.e(b1)=0.000321, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.008, 0.009). The C. I for B2 is approx. (-0.002, -0.001).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **94.6 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.946/2) / (0.054/130) = 1128.

You can calculate the probability of F >1128 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.536 and 3.361 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

VPD Analysis for Weyers Cave

Weyers Cave Dataset Head:

Table

Description automatically generated

Weyers Cave Dataset Description:

Table

Description automatically generated

Weyers Cave Rate of Change in VPD per Year

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

Weyers Cave Correlation Heat Map

Teams

Description automatically generated with low confidence

Linear Regression for Weyers Cave

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results

Graphical user interface, table

Description automatically generated

Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0164 and **humidity** decreases by 0.0063. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.0004991**

**standard error(humidity) = 0.0003902**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0164 and se(b1) = 0.0004991, So,

t = (0.0164 – 0) / 0.0004991 = **32.857**

Similarly, b2= -0.0063, se(b2) = 0.0003902

                   t = (-0.0063 – 0) / 0.0003902 = **-16.144**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (68.054 and -41.45) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0164, s.e(b1)=0.0004991, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.015, 0.017). The C. I for B2 is approx. (-0.007, -0.005).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **96.7 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.967/2) / (0.033/130) = 1889.

You can calculate the probability of F >1889 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.536 and 3.361 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

VPD Analysis for Philadelphia

Philadelphia Dataset Head:

Table

Description automatically generated

Philadelphia Dataset Description:

Table

Description automatically generated

Philadelphia Rate of Change in VPD per Year

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

Chart, line chart

Description automatically generated

Philadelphia Correlation Heat Map

A picture containing graphical user interface

Description automatically generated

Linear Regression for Philadelphia

Chart, line chart

Description automatically generated

Interpreting OLS Regression Results

Graphical user interface

Description automatically generated

Dependent Variable: Dependent variable is one that is going to depend on other variables. In this regression analysis **Y** is our dependent variable because we want to analyze the effect of **X**on **Y.** Here the Dependent Variable is **VPD**.

**Model:**The method of **Ordinary Least Squares (OLS)** is the most widely used model due to its efficiency. This model gives the best approximate of true population regression line. The principle of OLS is to minimize the square of errors (∑**ei2)**.

**Number of observations:**The number of observations is the size of our sample, i.e., N = 132.

**Degree of freedom(df) of residuals:**

Degree of freedom is the number of independent observations based on which the sum of squares is calculated.

D.f Residuals = 132 – (1+1) = 130

Degree of freedom (D.f) is calculated as,

 Degrees of freedom, D***. f = N – K***

*Where,****N =****sample size (no. of observations) and K* ***=****number of variables + 1*

**Df of model:**

***Df of model = K – 1 = 2 – 1 = 1,***

*Where,****K =****number of variables + 1*

**Coefficient term:**The coefficient term tells the change in **Y**for a unit change in **X i.e** if **VPD**rises by 1 unit, then **avgtempF**rises by 0.0213 and **humidity** decreases by 0.0096. If you are familiar with derivatives, then you can relate it as the rate of change of **Y**with respect to **X.**

**Standard error of parameters:**Standard error is also called the standard deviation. Standard error shows the sampling variability of these parameters.

**standard error(avgtempF) = 0.0007514**

**standard error(humidity) = 0.000631**

**t – statistics:**

In theory, we assume that error term follows the normal distribution and because of this the parameters **avgtempF and humidity**also have normal distributions with variance calculated in above section. Let avgtempF and humidity be b1 and b2 respectively.

 That is,

* **b1 ∼ N (B1,**σ**b12)**
* **b2**∼ **N (B2,**σ**b22)**

Here **B1and B2 are** true means of b1 and b2.

t – statistics are calculated by assuming following hypothesis –

* ***H0: B1 = 0 (avgtempF has no influence on VPD)***
* ***Ha: B1 ≠ 0 (avgtempF has significant impact on VPD)***
* ***H0: B2 = 0 (humidity has no influence on VPD)***
* ***Ha: B2 ≠ 0 (humidity has significant impact on VPD)***

Calculations for t – statistics:

**t = (b1– B1) / s.e (b1)**

From summary table, b1= 0.0213 and se(b1) = 0.0007514, So,

t = (0.0213 – 0) / 0.0007514 = **28.347**

Similarly, b2= -0.0169, se(b2) = 0.000631

                   t = (-0.0096 – 0) / 0.000631 = **-15.212**

**p – values:**

In theory, we read that p-value is the probability of obtaining the t statistics at least as contradictory to H0 as calculated from assuming that the null hypothesis is true. In the summary table, we can see that the P-value for both parameters is equal to 0. This is not exactly 0, but since we have much larger statistics (28.347 and -15.212) p-value will be approximately 0.

If you know about significance levels, then you can see that we can reject the null hypothesis at almost every significance level.

**Confidence intervals:**

There are many approaches to test the hypothesis, including the p-value approach mentioned above. The confidence interval approach is one of them. 5% is the standard significance level (∝) at which C. I’s are made.

C.I for B1is **(b1– t∝/2 s.e(b1), b1+ t∝/2 s.e(b1))**

Since ∝ = 5 %, b1= 0.0213, s.e(b1)=0.0007514, from t table, t0.025,130= **1.97838**,

After putting values, the C.I for B1is approx. (0.02, 0.023). The C. I for B2 is approx. (-0.011, -0.008).

While calculating p values we rejected the null hypothesis we can see same in C.I as well. Since 0 does not lie in any of the intervals so we will reject the null hypothesis.

**R – squared value:**

R2is the coefficient of determination that tells us how much percentage variation independent variable can be explained by independent variable. Here, **94.2 %** variation in Y can be explained by X. The maximum possible value of R2 can be 1, meaning the larger the R2 value, the better the regression.

**F – statistic:**

F test tells the goodness of fit of a regression. The test is like the t-test or other tests we do for the hypothesis.

Inserting the values of R2, n and k, F = (0.942/2) / (0.058/130) = 1060.

You can calculate the probability of F >1060 for 1 and 130 df, which comes to approx. 0. From this, we again reject the null hypothesis stated above.

The remaining terms are not often used. Terms like Skewness and Kurtosis talk about the distribution of data. Skewness and kurtosis for the normal distribution are 0.499 and 3.648 respectively. The Jarque-Bera test is used for checking whether an error has a normal distribution or not.

Code Explanation: (Only Explaining for Palo Alto City Cause Code is similar for other Cities too, The Whole Project is Done in Python Language, Jupyter Notebook)

Data Extraction:

Table

Description automatically generated with low confidence

This is Data Extraction Part where we Read the Data from an Excel File

Graphical user interface, text

Description automatically generated

Over here we have only extracted Important Features required to build the Model

Table

Description automatically generated

We have used VPD Formula, Inserted VPD as an Extra Feature in the Dataset for the Model

Data Description:

Table

Description automatically generated

This is a Brief Summary of our Data. Over here we can see all the Ranges of Quantiles for VPD score. Mean and the Median for VPD score. Min Max and Standard Deviation for VPD score.

Data Visualization:

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Graphical user interface, text, application, email

Description automatically generated

These are the Yearly Averages for VPD for all Years Data

Graphical user interface, application, Word

Description automatically generated

We have now Displayed Yearly Averages for VPD for all Years Data

Chart, scatter chart

Description automatically generated

This is the Rate of Change of VPD per Year which shows a Line Plot on each Scattered Values Present in Dataset

Chart, box and whisker chart

Description automatically generated

This is the Box Plot for the Dataset

Chart

Description automatically generated

This is a Correlation Heat Map where we have Displayed the Correlation between All Important Features

Chart, line chart

Description automatically generated

This is a Relational Plot which indicates the Rate of Change of VPD per Year for Specific Months where Lighter Shades are earlier Months of the Year and Darker Shades are Later Months of the Year

Data Analytics Using Linear Regression Model:

Graphical user interface, text, application, email

Description automatically generated

Here, we have used VPD as our Target or Dependent Variable and Temperature and Humidity as our Independent Variables. In short, we are going to use Temperature and Humidity Features to Predict VPD Outcomes in the Linear Model. Over Here we are going to use Linear Regression Model.

What is Liner Regression?

Linear regression analysis is used **to predict the value of a variable based on the value of another variable**.

Why do we use it?

To Determine the Strength of the Predictor Variable (e.g., VPD in this case scenario)

Where do we use it?

Forecasting an event or Trend Forecasting (e.g., Weather Forecasting in this case scenario)

A picture containing graphical user interface

Description automatically generated

Table

Description automatically generated with low confidence

A picture containing graphical user interface

Description automatically generated

Linear Regression Formula:

y = m \* x + c

where, y is the Outcome Variable

x is the Independent Variable

m is the Slope

c is the y Intercept

How Linear Regression Model is used in our Project?

We have two Independent Variables Temperature and Humidity, and we have One Outcome Variable VPD. So, the Formula is,

y = m1 \* x1 + m2 \* x2

where, y is the VPD prediction

x1 is Temperature Variable

x2 is Humidity Variable

m1 is Temperature Coefficient

m2 is Humidity Coefficient

Verification of the Model:

Given: x1 = 75

x2 = 60

m1 = 0.0289

m2 = -0.0169

c = 0

To Prove: y = 1.154

Proof:

y = m1 \* x1 + m2 \* x2

= 75 \* 0.0289 + 60 \* (-0.0169)

= 2.1675 – 1.014

= **1.154**

Note: x1 and x2 are Values obtained from x\_test table

m1 and m2 are Values obtained from OLS Regression Table

y is obtained from prediction table

Chart, line chart

Description automatically generated

This is the Quantile Plot of the Model. Over here we can Justify 99.3% accuracy and prove that there is Normal Distribution for this Model.

Results:

There is a Linear Growth in VPD every year for Palo Alto and Philadelphia.

This means that, Temperatures are rising every year in these 2 cities and Humidity is decreasing Significantly.

In the above Pages or Slides we Proved that Global Warming is Linearly Increasing every year in these 2 Cities.

We do not see any Significant Change in VPD for other Cities.

What did I Learn?

I Learned what is it meant to do a Research Project and Dive Deep into every aspect or Features that are Important.

I Learned that I don’t know a lot of stuff in my own field and realized that I have a lot of scope to Learn and Practice for the rest of my Life.

I learned that we cannot neglect Global Warming Trends and we should care about Climate Change.

I realized that VPD is a very Significant Factor to Forecast Global Warming Trends.

My Data Visualization Skills improved in this Project.

I learned that the ability to explain your Project by giving simple analogy to 15 years old is Far Greater than What Skillsets you Possess.

Was I Successful?

I would say Partially Successful.

Yes, I was Successful in getting an Output.

I was Successful in showing a trend and Proving my Point.

But this Project felt like an Incomplete Project.

I hardly got 2 months to execute this project.

What I did was Normal Weather Forecasting, but I wanted to do something more beyond that.

My purpose in this Project was to show what factors affect Suicide in Professional Athletes, but I was able to work on only one factor viz, VPD or Global Warming Factor but couldn’t involve other Important Factors that do cause Suicides.

If I didn’t have any Time Constraints, I would have tried to add a minimum of 5-6 more features that cause Suicides.