**SDS-Mini-Project**

**Sem-3**

**Group members:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No.** | **Name** | **SRN** | **Section** |
| 1. | K Manish Gowd | PES1UG19CS200 | C |
| 2. | Justin James | PES1UG19CS198 | C |
| 3. | Harshit Khashoo |  | C |
| 4. | Dhanush Patel YP |  | C |

**Report**

**1. Abstract:**

This dataset consists of the weather records for a month taken with the observations being taken down and noted per hour. It consists of various aspects of weather like precipitation, humidity, wind speed, wind bearing, visibility, loud cover, pressure, temperature, apparent temperature against a particular instant of time. The summary and the daily summary are noted down appropriately as per the parameters collected. Most of the parameters are numerical with the remaining few being categorical. The objective of this project is to make optimum use of the dataset to extract maximum and most relevant information out of the dataset for various purposes. The first section of the project would consist of descriptive statistics while the second half would contain inferential statistics. With the help of the relationships generated between various variables, we can appropriately put this to a great use by inferring various parameters from it. While this is just a sample of 720 records, we can easily put this to a great use to infer about various weather parameters on the other given days with optimal accuracy and precision along with minimal margin of error.

**2. Introduction:**

This dataset consists of the weather records for a month taken with the observations being taken down and noted per hour. It consists of various aspects of weather like precipitation, humidity, wind speed, wind bearing, visibility, loud cover, pressure, temperature, apparent temperature against a particular instant of time. The summary and the daily summary are noted down appropriately as per the parameters collected.

Most of the parameters are numerical with the remaining few being categorical. The objective of this project is to make optimum use of the dataset to extract maximum and most relevant information out of the dataset for various purposes. The first section of the project would consist of descriptive statistics while the second half would contain inferential statistics. With the help of the relationships generated between various variables, we can appropriately put this to a great use by inferring various parameters from it. While this is just a sample of 720 records, we can easily put this to a great use to infer about various weather parameters on the other given days with optimal accuracy and precision along with minimal margin of error.

**3. Dataset:**

This dataset consists of the weather records for a month taken with the observations being taken down and noted per hour. It consists of various aspects of weather like precipitation, humidity, wind speed, wind bearing, visibility, loud cover, pressure, temperature, apparent temperature against a particular instant of time. The summary and the daily summary are noted down appropriately as per the parameters collected.

Most of the parameters are numerical with the remaining few being categorical. The objective of this project is to make optimum use of the dataset to extract maximum and most relevant information out of the dataset for various purposes. The first section of the project would consist of descriptive statistics while the second half would contain inferential statistics. With the help of the relationships generated between various variables, we can appropriately put this to a great use by inferring various parameters from it. While this is just a sample of 720 records, we can easily put this to a great use to infer about various weather parameters on the other given days with optimal accuracy and precision along with minimal margin of error.

**4. Preprocessing or Data Cleaning:**

Techniques used:

* Correction of inconsistent capitalization
* Dropping of excessively repetitive parameters and observations
* Replacement of NaN with 0
* Replacing NaN by techniques such as imputing by interpolation, etc.

**5. Exploratory Data Analysis: What are the insights obtained from EDA. Represent**

**them in terms of various graphs along with explanation.**

Various methods of representing data:

**6. Hypothesis Testing:**

## Prerequisites:

* We need to convert the temperature to K since that is the SI unit. Therefore add 273.15 to each unit of temperature.
* We also need to convert the pressure to pascal since that's it's SI unit.  
  Since the pressure is in millibars, convert it in terms of bars by dividing by 1000.  
  We know that 1 bar = 10,000 Pa.  
  Therefore multiply by the pressure in bar by 10,000 to obtain the pressure in terms of it's SI units.

We know that for a closed system, temperature is directly proportional to pressure.

But the graph for the sample dataset doesn't seem to depict it. It rather approximates an inverse relationship between pressure and temperature.

### Can we prove that there is no apparent relationship between temperature and pressure?

The general assumption here is that the temperature and pressure are directly related

### Therefore:

**H0**: The hypothesis that sample observations result purely from chance. (T is directly proportional to P, and our observations have just been resulted by chance)  
  
**H1**: The hypothesis that sample observations are influenced by some non-random cause. (T isn't directly proportional to P)

## b. Statistical tests

H0: **T α P** is **True**  
  
H1: **T α P** is **False**

Since the sample size is appreciably large enough, we can approximate it to a sample from a normal distribution. Also, since the standard deviation is also known, it's safe to say that we can use the z-table in order to perform the statistical tests.

Since the general belief is that temperature and pressure are directly related to each other, i.e. **(T α P)**

Therefore, the following equation would be true:

**(T = kP)** (where k is an arbitrary constant)

**T/P = k**

Therefore, we have to ensure that the value of k remains as constant as possible with minimum possible deviation and negligible error. It's also crucial to avoid any possible bias in order to ensure proper results.

We know that the average of the ratio of standard temperature(273.15 K) and standard pressure(10,000 bar) is a constant i.e. 0.0027315

Using z-score to standardize the data: We obtained the mean of the ratio as 0.002868 We obtained the standard deviation of the ratio as 6.569896069161947e-05

Therefore,

x=0.002868  
μ=0.0027315  
σ = 6.569896069161947e-05

Therefore, the z-score is as follows:

**z = ( x - μ ) / σ**

Upon, solving, we obtain, P-value=0.038694142721681285

### 5% statistical significance:

Since P-Value is less than 0.05, (taking at 5% statistical significance), we reject H0, the original hypothesis, and accept our claim that there is no specific relationship between temperature and pressure to be plausible

But, we know that pressure is directly proportional to temperature only at constant volume. We also know that the volume doesn't always remain constant. The change in volume to most extent would be practically significant. But the changes at times are definitely statistically significant. That's the main reason why we the null hypothesis was rejected by taking 5% statistical significance.

Therefore, a better solution is to increase the level of accuracy and bring the level of statistical significance down to 1%, which obviously seems to be a better choice in this scenario.

### 1% statistical significance:

So, since P-value is greater than 0.01, (taking at 1% statistical significance), we do not reject H0, the original hypothesis, and continue with the accepted fact i.e. the null hypothesis

### Therefore, we can confidently conclude that the null hypothesis is supported.

**7. Results and Discussion of your problem statement**

**With the assistance of data and results proved above, we obtain various inferences on the nature of the dataset as mentioned below:**

Most of the days were rainy and cloudy in nature since the observations in the dataset have been from the month of September, during which monsoon/rainy season is prevalent in most parts of the Northern Hemisphere in and around the Tropic of Cancer. We can also easily infer and predict the weather in the forthcoming couple of months, using the given statistics to find the parameters. We can also predict the weather in the similar months in the upcoming few years, but at the same time, we would also need to take into account the rising rate of various environmental effects like greenhouse effect leading to global warming thereby leading to imbalanced seasons which result in excess droughts and floods. The average temperature of the earth is also rising slowly but steadily. So, it wouldn't be wrong to say that the inferences and predictions made in the future i.e. after 5 years and more... wouldn't be plausible as they are for the near future, thereby leading to an increase in the degree of uncertainty and many more assumptions and null hypotheses being rejected. But, we can predict the weather to some extent in general even in the upcoming years, just that the degree of uncertainty would be witnessing an exponential rise, which isn't of any help to the statisticians even by the slightest of means.

So, take care of the environment as well :)