**TERM Project DSC530-T302**

**EDA Summery Document:**

Prashant Raghuwanshi

**Bellevue University**

Guided by Professor  Matthew Metzger

**EDA**

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

**Statistical/Hypothetical Question**

* How accurately can we classify whether a sample from a body of water is potable, given its chemical and physical characteristics?
  + Explore more on variables and list down the information about them
  + Do the used data set is having missing values
  + Do the variables are having any unusual observations (outliers) or any gaps in the data.
  + Find out the type of distributions in datasets variables
  + Do any variables are having normal distributions
  + What CDF and PMF says if compare the portable and non portable datasets for Ph variables
  + Find out the Analytical distribution for ph variable (Actual Vs Model) data
  + What hypothesis test results says
  + How portability is fit linearly with other variables like Ph and Hardness

**Outcome of EDA**

Acquired very good EDA knowledge and able to relate the use case of Stats & EDA in currently working data project.

Outcome of the research question:

1. Fixed the missing values in dataset
2. Able to find out below dependant and Predictor variables from the dataset

Predictor variable 🡪 Portable

Dependant variable 🡪 Ph , Hardness

1. Able to remove outliers from datasets
2. Not all variables values are correlated with the predictor variables
3. However by using some of the variables like Ph , Hardness we can easily predict the portability of given water sample
4. CDF shows the Ph values of Portable water is slight less as compared to the Ph value of non portable water

What do you feel was missed during the analysis?

1. In my analytical distribution comparison, when I have passed less record count (then actual) and for model data I took more data, I am getting normal difference in Ph value.

However when I am using actual sample data records count and this difference is getting increase abnormally (slide 17)

1. I have performed fit regression analysis on the top of clean dataset (after removing outliers), seems p values are not coming as expected.

Were there any variables you felt could have helped in the analysis?

Predictor variable 🡪 Portable

Dependant variable 🡪 Ph , Hardness

Were there any assumptions made you felt were incorrect?

While creating the different stats visualization plots, most of the places I have used rounded value of Ph values , some places it make plot smooth , however some places it impacted the results

What challenges did you face, what did you not fully understand?

1. Getting required sample data is one of the challenging task
2. Want to explore more on regression analysis tables & examples

REFERENCES

* Think Stats by Allen B Downey
* Web references:
* <https://courses.lumenlearning.com/introstats1/chapter/skewness-and-the-mean-median-and-mode/>
* https://www.kaggle.com