https://www.pes.edu/wp-content/uploads/2019/09/pes_logo.png

**B.TECH. (CSE)**

**III SEMESTER**

**UE19CS203 – STATISTICS FOR DATA SCIENCE**

**PROJECT REPORT**

**ON**

“ANALYSIS OF KEPLER EXOPLANET SEARCH RESULTS DATASET”

SUBMITTED BY

**NAME SRN SECTION**

1. **AKASH V PES2UG19CS023 A**
2. **ANURAG R SIMHA PES2UG19CS052 A**
3. **ACHYUT JAGINI PES2UG19CS013 A**
4. **ACHYUTA B MUDHOL PES2UG19CS014 A**

**AUGUST – DECEMBER 2020**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**ELECTRONIC CITY CAMPUS,**

**BENGALURU – 560100, KARNATAKA, INDIA**

**ABSTRACT OF THE PROJECT:**

The purpose of this study is to use statistical analysis to confirm the presence of an exoplanet using a Kepler Telescope. The data set being studied contains parameters regarding the exoplanet found in the galaxy. We cleaned the dataset by fixing typos and handling missing data by imputing. We obtain various useful insights into how various factors such as the relation between declination and equatorial temperature. We also test the hypothesis(claim) that mean of insolation temperature is equal to population mean and derive a result that prompts us to reject this hypothesis.

**INTRODUCTION:**

The purpose of this study is to use statistical analysis to confirm the presence of an exoplanet using a Kepler Telescope. The Kepler Space Observatory is a NASA-build satellite which was launched in 2009. The telescope is dedicated to searching for exoplanets in star systems besides our own, with the goal of possibly finding other habitable planets besides our own. The original mission ended in 2013 due to mechanical failures, but the telescope has nevertheless been functional since 2014 on a "K2" extended mission. Kepler had verified 1284 new exoplanets as of May 2016. As of October 2017, there are over 3000 confirmed exoplanets total (using all detection methods, including ground-based ones). The telescope is still active and continues to collect new data on its extended mission.

**Problem Statement:** To analyze the Kepler Exoplanet search results and analyze the parameters of the exoplanets by cleaning the dataset, obtaining exploratory data analysis, claiming certain hypothesis as null hypothesis and deriving a certain conclusion from all of the above.

**DATASET:**

This dataset was taken from [www.kaggle.com](http://www.kaggle.com) , a website dedicated to collecting and comprehending statistical datasets It allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

The dataset contains 4 categorical variables namely, kepoi\_name, kepler\_name, koi\_disposition and koi\_pdisposition.

It contains the following 7 discrete numerical variables: koi\_fpflag\_nt, koi\_fpflag\_ss, koi\_fpflag\_co, koi\_fpflag\_ec, koi\_steff, koi\_steff\_err1, koi\_steff\_err2.

It also contains the following continuous variables: plnt\_disp\_confidence,flag\_nTransitLk,flag\_scndEvent,flag\_centroidOffset,flag\_ephMatch,orbital\_period,transit\_epoch,impact\_parameter,transit\_duration,transit\_depth,planetary\_radius,equ\_temp,insolation\_flux,transit\_sigToNoise,stellar\_eff\_temp,stellar\_surf\_gravity,stellar\_radius,right\_acension,declination,kepler\_magnitude.

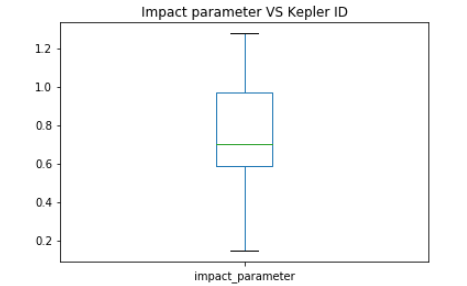
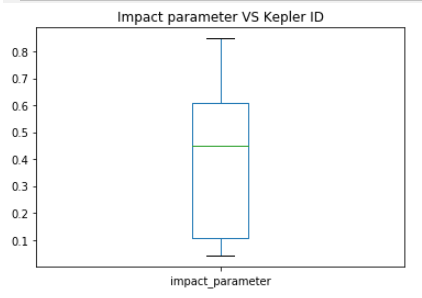
**PREPROCESSING AND DATA CLEANING:**

On testing for null values, we found out that there was a total of 40557 missing values in the dataset, across various discrete and continuous numerical columns. We rectified the missing data by dropping the specific rows and columns which have the maximum missing values.

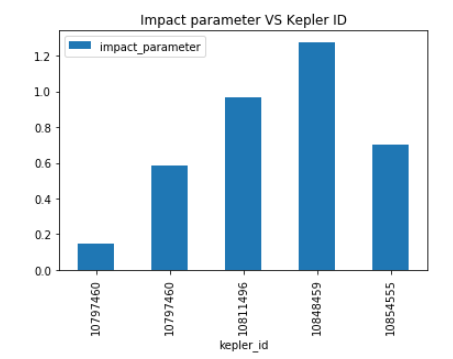
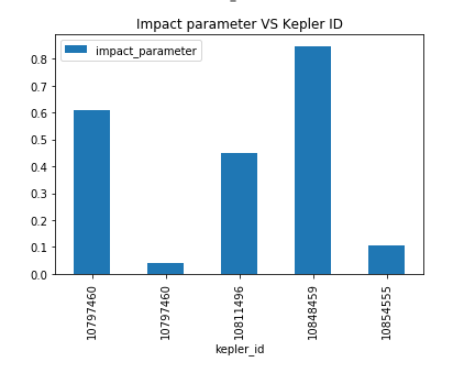
**koi\_pdisposition**.

* + This column describes whether a KOI [Kepler Object of Interest] is considered a candidate for an exoplanet, or a false positive, or a confirmed exoplanet.)
  + **koi\_disposition** is the same, but also incorporates newer, peer-reviewed/published information from the Exoplanet Archive, so keep that and get rid of **koi\_pdisposition,** which is from the Kepler pipeline.
* **koi\_tce\_plnt\_num** and **koi\_tce\_delivname.**
  + **koi\_tce\_plnt\_num** is the TCE (Threshold-Crossing Event) Planet Number federated to the KOI.
  + **koi\_tce\_delivname** is TCE delivery name corresponding to the TCE data federated to the KOI.
* all the error columns (**koi\_period\_err1**, **koi\_impact\_err2**, etc...)
  + These columns contain uncertainty values for their corresponding columns, e.g. **koi\_period\_err1** and **koi\_period\_err2** are the positive and negative uncertainties for koi\_period.
* **kepler\_name.**
  + Kepler number name in the form "Kepler-N," plus a lower-case letter, identifying the planet.
  + Only the confirmed planets have these, everything else has a NaN value here. when we try to drop rows with NaN values, everything besides the confirmed exoplanets will get dropped... which we don't want. We can always cross-reference from the cleaned-out dataset back to this one if we really want the object's Kepler name. Hence, data is dropped.

**EXPLORATORY DATA ANALYSIS**

**** ****

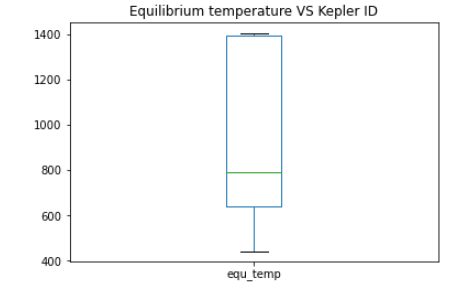
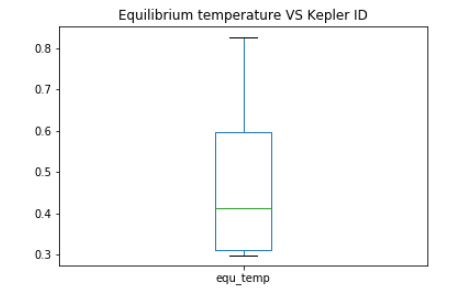
**Plot based on the values in the dataset**  **Normalized plot**

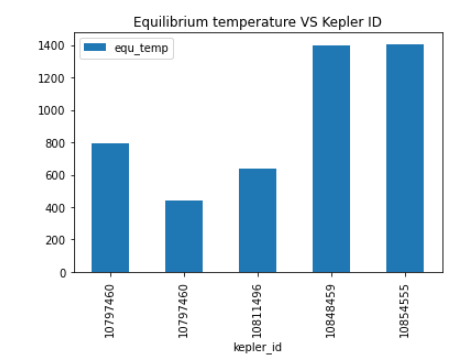
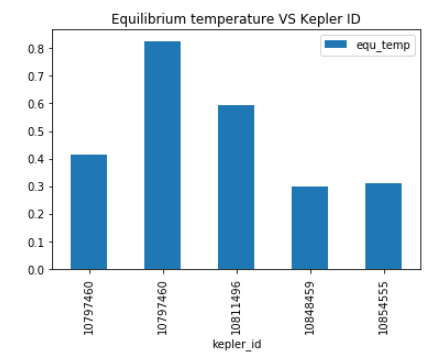
**Plot based on the values in the dataset Normalized plot**

The above graphs are plotted between Impact Parameter and Kepler Id. We are comparing the first five kepler ids and its Impact parameter. We can say that id 10848459 has the highest impact parameter and kepler id 10797460 has the least impact parameter. Impact parameter infers the radius of planet and stellar radius so higher the planetary radius higher the impact

parameter.

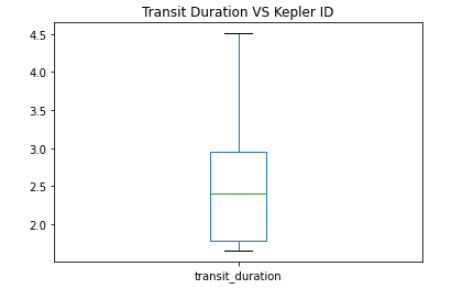
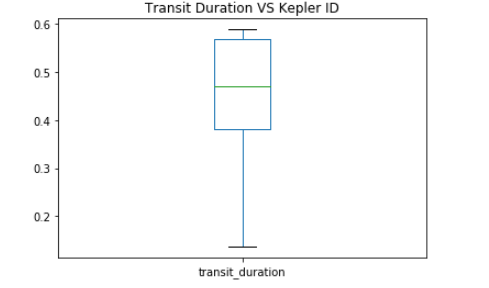
 

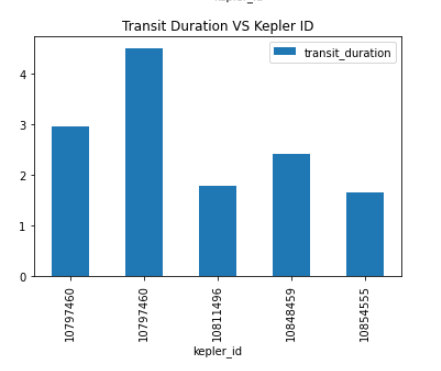
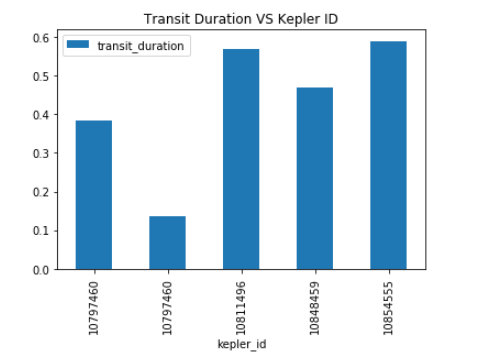
**Plot based on the values in the dataset**  **Normalized plot**

** **

**Plot based on the values in the dataset Normalized plot**

The above graphs are plotted between equilibrium temperatures and kepler ids. We are comparing the first five kepler ids and its Equilibrium Temperature. We can say that id 10854555 has the highest Equilibrium Temperature and kepler id 10797460 has the least. Equilibrium Temperature infers an approximate temperature of the planet.

The above graphs are plotted between Transit Duration and kepler ids. We are comparing the first five kepler ids and its Transition Duration. We can say that id 10811496 has the least Equilibrium Temperature and kepler id 10797460 has the highest. Transit Duration refers to the duration of the observed transits. Duration is measured from first contact between the planet and star until last contact.

**HYPOTHESIS TESTING:**

1. I make a claim (hypothesis) that population mean of the sample is the sample mean of the declination. So, we perform one sample t-test where we check the mean of declination column is the same as the population mean. So, we get p value which is less than 0.05(alpha) hence rejects the null hypothesis.
2. I make another claim (hypothesis) that population mean of mean of orbital period of planet dispositioned as CONFIRMED and mean of orbital period of planet dispositioned as FALSE POSITIVE are different and mean of orbital period of planet dispositioned as CONFIRMED and mean of orbital period of planet dispositioned as FALSE POSITIVE are same. We get p value less than 0.05(alpha) which rejects the null hypothesis.
3. Finally, I claim that population mean is equal to sample mean of the insolation flux. On conducting the hypothesis testing we get z value less than 1.96 which is the critical z value, hence rejecting the null hypothesis.

**RESULTS:**

This project was successful in pinpointing useful insights into the various search results of exoplanets using Kepler’s telescope. We have successfully cleaned the dataset by fixing typos and handling missing data by imputing. We have obtained various useful insights into how various factors such as the relation between declination and equatorial temperature. We also tested the hypothesis(claim) that mean of insolation temperature is equal to population mean and derived a result that prompts us to reject this hypothesis.

In conclusion, with the help of the powerful tools than python provides, we have successfully analyzed and obtained meaningful insights for the Kepler Exoplanet search results dataset.