Project Report

Data Analysis on NASA Exoplanet Archive

Arnab Kumar Chand (CWID - 20012426)

Bharath Beeravelly (CWID - 20015625)

Ratan Chowdhary (CWID - 20011581)

1 Introduction

In this project, we delve into the intriguing realm of exoplanet analysis, utilizing a meticulously chosen dataset from the NASA Exoplanet Archive. Our study focuses on a subset of data consisting of four pivotal columns: koi_teq, representing the equilibrium temperature of exoplanets; koi_steff, detailing the effective temperature of the host stars; koi_fpflag_ss, indicating false positives due to solar system events; and koi_fpflag_co, denoting false positives caused by contamination in the optical region. By selecting these specific parameters, our objective is to explore and decipher complex relationships in exoplanetary systems. This project aims to provide a deeper understanding of the characteristics and dynamics of exoplanets and their host stars, leveraging advanced statistical methods and modeling techniques.

2 Data Exploration

2.1 Meet the Data

The NASA Exoplanet Archive is a comprehensive database containing information about exoplanets—planets that orbit stars outside our solar system—along with details about their host stars. This repository primarily gathers data from observations made by NASA's Kepler space telescope, which has been instrumental in detecting and confirming numerous exoplanets.

From a pool of 49 columns, 22 were identified as Kepler error-replated columns. Prioritizing statistically advantageous variables, the search for normally distributed quantitative features revealed a lack of precise matches. Instead, through visual scrutiny, columns exhibiting semblances of normal distributions were discerned, offering potential utility despite not perfectly meeting the normality criterion. Concurrently, attention was drawn to qualitative data showcasing binary representations within the dataset.

Four variables have been meticulously chosen from the dataset for analysis.

Variable	Data type	Label
koi_teq	Continuous	Equilibrium temperature of planet (Kelvin)
koi_steff	Continuous	Effective temperature of host star (Kelvin)
koi_fpflag_ss	Categorical	Stellar Eclipse Flag
koi_fpflag_co	Categorical	Centroid Offset Flag

Table 1 Information about selected Quantitative and Qualitative columns

Qualitatively, the dataset embraces dichotomous aspects. On the quantitative front, the dataset contains continuous variables. The dataset appears unprocessed, possibly harboring missing values, and outliers.

2.2 Data Cleaning

In the pursuit of refining the dataset and ensuring its reliability, a thorough data cleaning process was executed. Columns afflicted with missing values were systematically removed from the dataset. This step was crucial to maintain the dataset's consistency and integrity for subsequent analyses. Outliers, which could potentially distort analysis outcomes, were identified and subsequently eliminated leveraging quartile-based methods. By establishing thresholds based on quartiles, this technique effectively culled extreme values, fostering a more robust and dependable dataset. Data points falling below **Q1 - 1.5 * IQR** or above **Q3 + 1.5 * IQR** were considered outliers.

2.3 Data Visualization

2.3.1 Quantitative columns

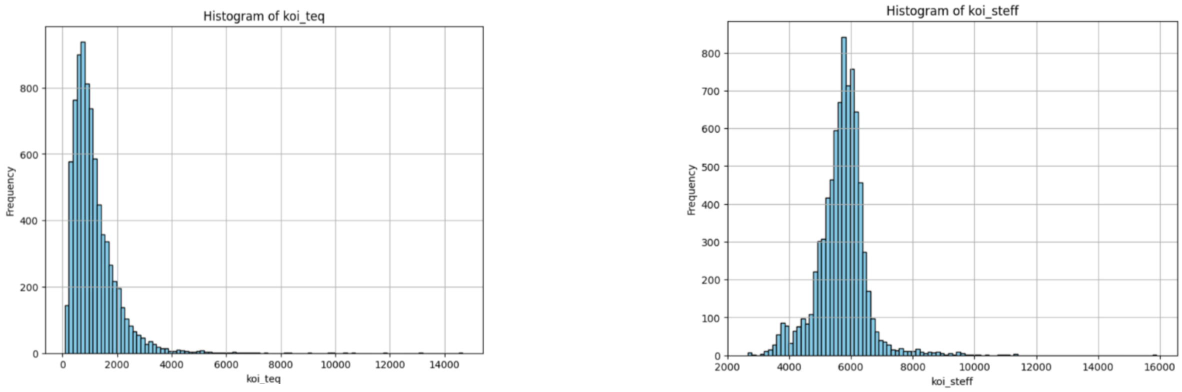

Figure 1: Histogram charts of columns koi_teq and koi_steff

The variables **koi_teq** and **koi_steff** represent continuous data points in the dataset. These columns contain outliers that need to be cleaned.

2.3.2 Qualitative columns

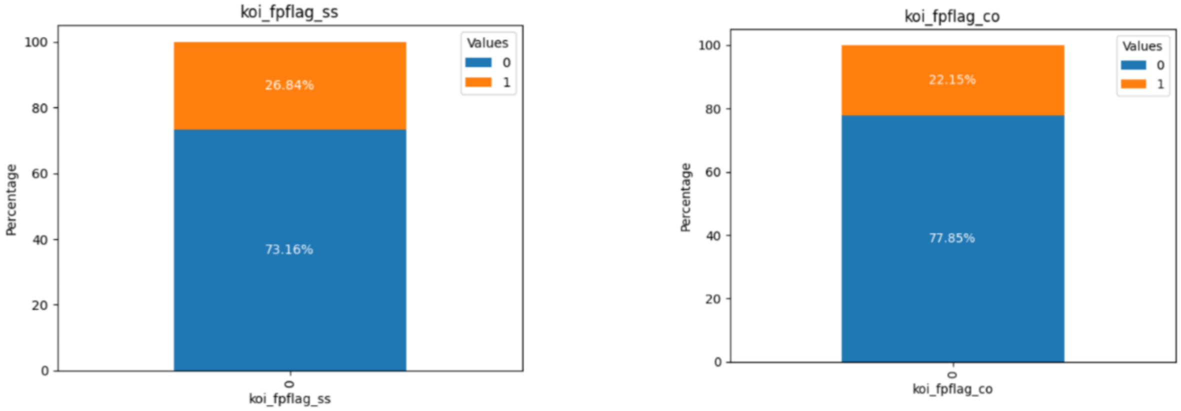

Figure 2: Stacked bar charts of columns koi_fpflag_ss and koi_fpflag_co

The stacked bar chart shows us that flags have unbalanced data.. For **koi_fpflag_ss**, approximately 73.16% of the dataset is labeled as 0, while 26.84% is labeled as 1. Similarly, for **koi_fpflag_co**, the distribution shows approximately 77.85% labeled as 0 and 22.15% labeled as 1. These percentages highlight the prevalence of each unique value within these specific flags.

	0’s	1’s
koi_fpflag_ss	73.16%	26.84%
koi_fpflag_co	77.85%	22.15%

Table 2: Distribution of 0 and 1 of qualitative columns

2.4 Summary Statistics

2.4.1 Quartiles

Quartiles divide a dataset into four equal parts. Q1 is the value below which 25% of the data fall. Q2 (median) is the value below which 50% of the data fall. Q3 is the value below which 75% of the data falls. The Interquartile Range is calculated as Q3 - Q1

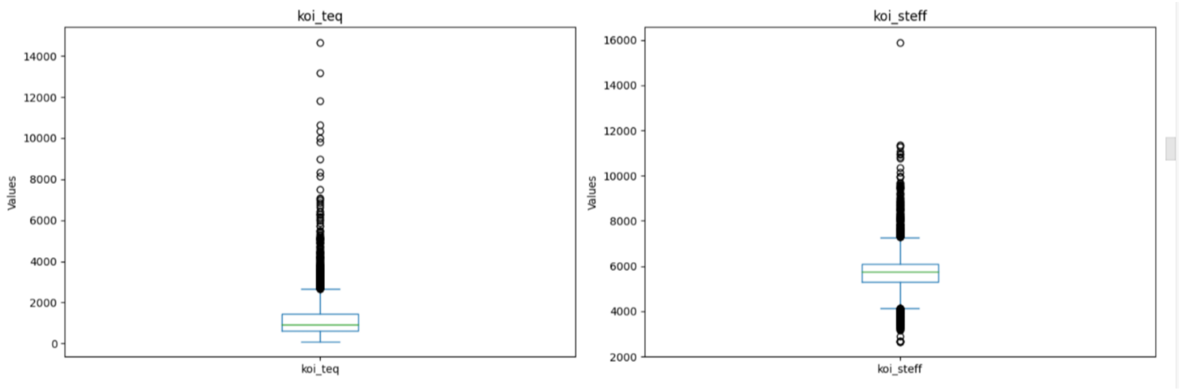

Figure 3: Box plot visualization of quantitative columns

From the box plot, we can interpret that koi_teq and koi_steff have outliers. It also shows us, koi_teq is right skewed and koi_steff have fat tails.

For **koi_teq**, the quartile values are outlined: Q1 (the first quartile) is 612.0, Q2 (the second quartile, which is also the median) is 938.0, and Q3 (the third quartile) is 1435.0. The interquartile range (IQR), representing the spread of the middle 50% of the data, is calculated as 823.0. Additionally, the lower bound for potential outliers is -622.5 and upper bound for potential outliers is 2669.5.

Concerning **koi_steff**, the quartiles are specified as Q1 = 5312.0, Q2 = 5761.0 (the median), and Q3 = 6098.0. The calculated interquartile range (IQR) is 755.75. The lower bound for potential outliers is 4185.625, while the upper bound is 7208.625.

	Q1	Q2	Q3	IQR	Lower Bound	Upper Bound
Koi_teq	612.0	938.0	1435.0	823.0	-622.5	2669.5
Koi_steff	5312.0	5761.0	6098.0	755.75	4185.625	7208.625

Table 3: Quartiles statistics

2.4.2 Mean

The mean, or average, is a measure of central tendency in a dataset. **Koi_teq**, representing the approximation for the temperature of the planet has a mean value of 1142.0410 Kelvin, while **Koi_steff**, denoting the photospheric temperature of the star, has a mean value of 5691.4291 Kelvin

	Koi_teq	Koi_steff
Mean	1142.0410	5691.4291

Table 4: Mean values of quantitative columns

2.4.3 Median

The median is the middle value in a dataset that separates the higher half from the lower half of the data when arranged in numerical order. **Koi_teq**, has a median value of 928.0 Kelvin, while **Koi_steff**, has a median value of 5767.0 Kelvin.

	Koi_teq	Koi_steff
Median	928.0	5767.0

Table 5: Median values of quantitative columns

2.4.4 Standard Deviation

The standard deviation is a statistic that measures the amount of variability or dispersion in a set of values. It represents the square root of the variance. **Koi_teq** has a standard deviation of 846.2481 Kelvin, while **Koi_steff** has a standard deviation of 788.2733 Kelvin.

	Koi_teq	**Koi_steff**
Standard Deviation	846.2481	788.2733

Table 6: Standard deviation of quantitative columns

2.5 Statistical Inference

2.5.1 Correlation analysis

We have used the Spearman correlation coefficient, as the data is not normally distributed. This is a non-parametric method, that does not assume the variables follow a specific distribution

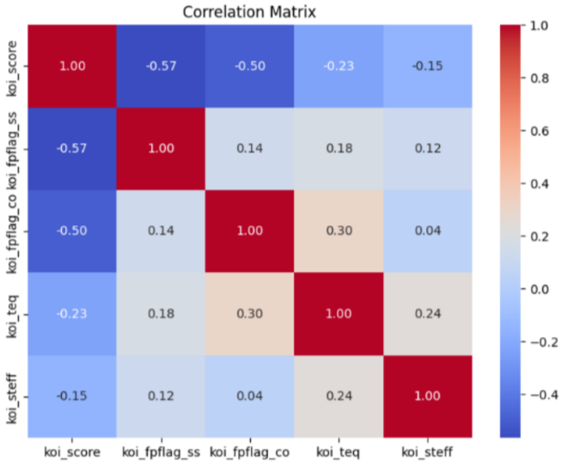

Figure 4: Correlation max of selected columns

The quantitative variables koi_teq and koi_steff have low correlation between them. The qualitative variables koi_fpflag_co and koi_fpflag_ss have low correlation as well. The correlation between quantitative and qualitative variables is either low or none.

koi_fpflag_ss is highly correlated with koi_score and has low correlation with koi_steff and koi_fpflag_co. **Koi_fpflag_co** is highly correlated with koi_score and has no correlation with koi_stef. **Koi_teq** has low correlation with all other columns. **Koi_steff** have no correlation with koi_fpflap_co and low correlation with other columns. **koi_score** is highly correlated with koi_fpflag_ss and have low correlation with koi_steff

2.6 Normal Distribution Assumptions

Before we started our analysis, we removed missing value rows and removed outliers using quartiles. We make assumptions through visualization, then we test them with statistical methods i.e QQ plots and Normality tests.

2.6.1 Histogram Plot:

Histograms are indispensable visual tools for understanding the distribution of data within columns, revealing key insights such as distribution shape—whether it's normal, skewed, or multimodal—along with measures of central tendency and variability. They swiftly illustrate the spread of values, highlight outliers, and depict the range covered by the dataset. Visually, koi_teq looks to be right skewed and koi_steff looks to be normally distributed. Figure 5, shows us, koi_teq is right skewed and not normal. Koi_steff have fat tails and looks normal

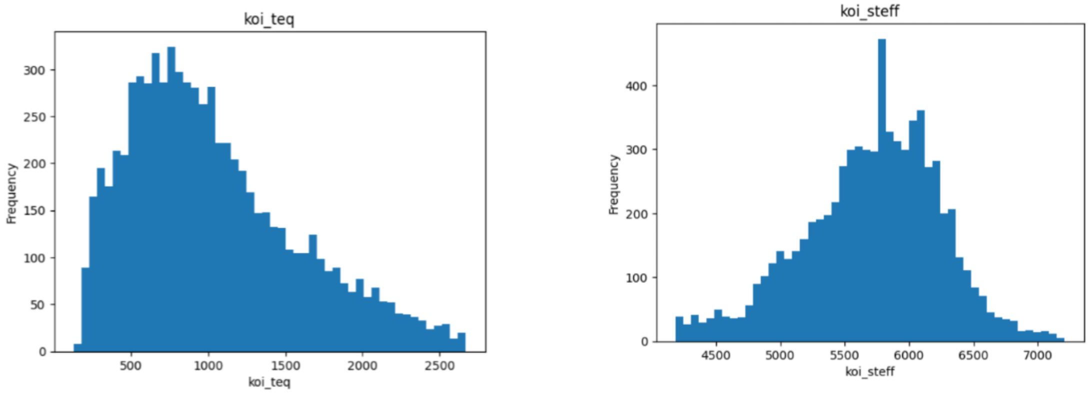

Figure 5: Histogram plot for quantitative columns

2.6.2 QQ Plots:

QQ plots (Quantile-Quantile plots) are graphical tools used to assess whether a given dataset follows a specific probability distribution, typically the normal distribution. QQ plots results show us that koi_teq is right skewed. And koi_steff is not normally distributed and has fat tails.

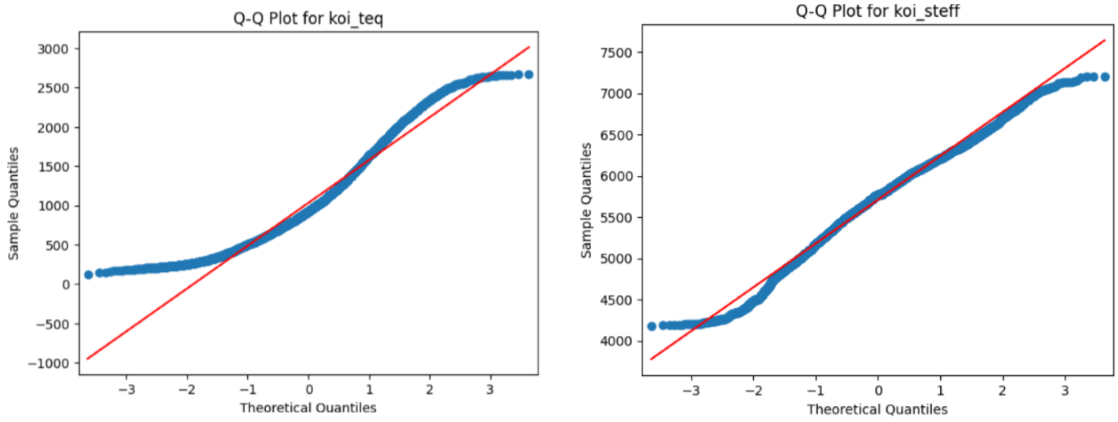

Figure 6: QQ plots for quantitative columns

2.6.3 Normality test:

The normality test we used is based on D'Agostino and Pearson's test that combines skew and kurtosis to produce an omnibus test of normality. The normality test yielded a statistic of 559.4202 and a p-value of 0.0000. The extremely low p-value (less than the common significance level of 0.05) indicates strong evidence against the null hypothesis (H0) that the data follows a normal distribution. Therefore, the conclusion is to reject the null hypothesis, indicating that the **koi_teq** dataset significantly deviates from a normal distribution. Similarly, the normality test for **koi_steff** resulted in a statistic of 154.2225 with a p-value of 0.0000. Again, the very low p-value suggests strong evidence against the assumption of normality for the koi_steff dataset. Hence, the null hypothesis of the data following a normal distribution is rejected.

The statistic value 559.4202 for koi_teq, explains the heavy tailedness and high skewness in the data. This aligns with QQ plot results. The koi_steff statistic is low i.e 154.2225, which again aligns with QQ plot results. It shows us, koi_steff have small fat tails and very less skewness.

	statistic	**p-value**
Koi_teq	559.4202	0.0000
Koi_steff	154.2225	0.0000

Table 7: Normality test statistic values and p-value

3 Univariate Analysis

Univariate analysis serves as a fundamental exploratory technique in statistical analysis, providing a comprehensive examination of individual variables in isolation. In this analytical approach, the focus is directed towards understanding the distribution, patterns, and characteristics of a single variable at a time, without considering the relationships with other variables. By scrutinizing the central tendencies, measures of dispersion, and the shape of the distribution, univariate analysis equips researchers and analysts with valuable insights into the inherent properties of a variable.

Within the realm of Univariate Analysis, our exploration delves into the application of the Central Limit Theorem (CLT) as a pivotal tool for obtaining more robust and representative samples from the population. Employing the CLT, we ascertain confidence intervals across diverse samples and varying confidence levels, providing a nuanced perspective on the precision of our estimations. Subsequently, our analytical journey extends to the realm of hypothesis testing, a powerful technique for extracting deeper insights and validating assumptions. Through a series of hypothesis tests, we aim to enhance our comprehension of the characteristics and behaviors inherent in the examined univariate data. Together, these analytical approaches not only refine our understanding of individual variables but also contribute to the broader statistical narrative, fostering a more informed interpretation of the underlying data structure.

3.1 Central Limit Theorem

The Central Limit Theorem (CLT) is a fundamental concept in statistics that plays a pivotal role in making statistical inferences about populations. It states that, regardless of the shape of the original population distribution, the distribution of the sample means will tend to follow a normal distribution as the sample size increases. This remarkable property allows statisticians to make reliable estimates and conduct hypothesis tests even when the underlying data may not be normally distributed. A key aspect of CLT is that the average of sample means and standard deviations will equal the population mean and standard deviation.

From all the available variables in the dataset, we have chosen the Equilibrium temperature as our population, on which we shall test CLT and draw our conclusions. Initially, we have decided that for each test case, the number of samples will remain the same, which is 500, so that there is no inherent bias induced in any test case. We have created multiple samples with varying sizes to observe the behavior of sample statistics. The sample sizes are 20, 50, 75 and 100. We have used simple random sampling to create these sample sizes.

In our analysis, we systematically computed the sample mean for each of the 500 samples, maintaining a consistent sample size throughout. The ensuing step involved visualizing these sample means through histograms, providing a graphical representation that facilitated the observation of the distribution's shape.

Our primary objective was to assess whether the observed distribution approximates the characteristics of a normal distribution. For this specific section, our focus was on leveraging visualizations alongside the aggregated results of the average sample means. Subsequent sections of our analysis will delve deeper into a more rigorous statistical examination, incorporating techniques such as confidence interval estimation and hypothesis testing. These approaches will allow us to rigorously scrutinize and validate the Central Limit Theorem.

Figure 1 presents a visual representation of the Central Limit Theorem (CLT) across varying sample sizes. Consistent with the CLT principle, as the sample size increases, the convergence of the sample means toward a normal distribution is evident. Notably, for a sample size of 20, the distribution exhibits a close approximation to normality, with a slight deviation observed in the rightmost histograms—a deviation unexpected in a truly normal distribution. However, as the sample sizes progress to 50, 75, and 100, these anomalies diminish, and the visualizations manifest a closer adherence to the characteristics of a normal distribution. The increasing sample sizes contribute to the progressive smoothing of the distribution curve, reinforcing the theoretical underpinnings of the CLT.

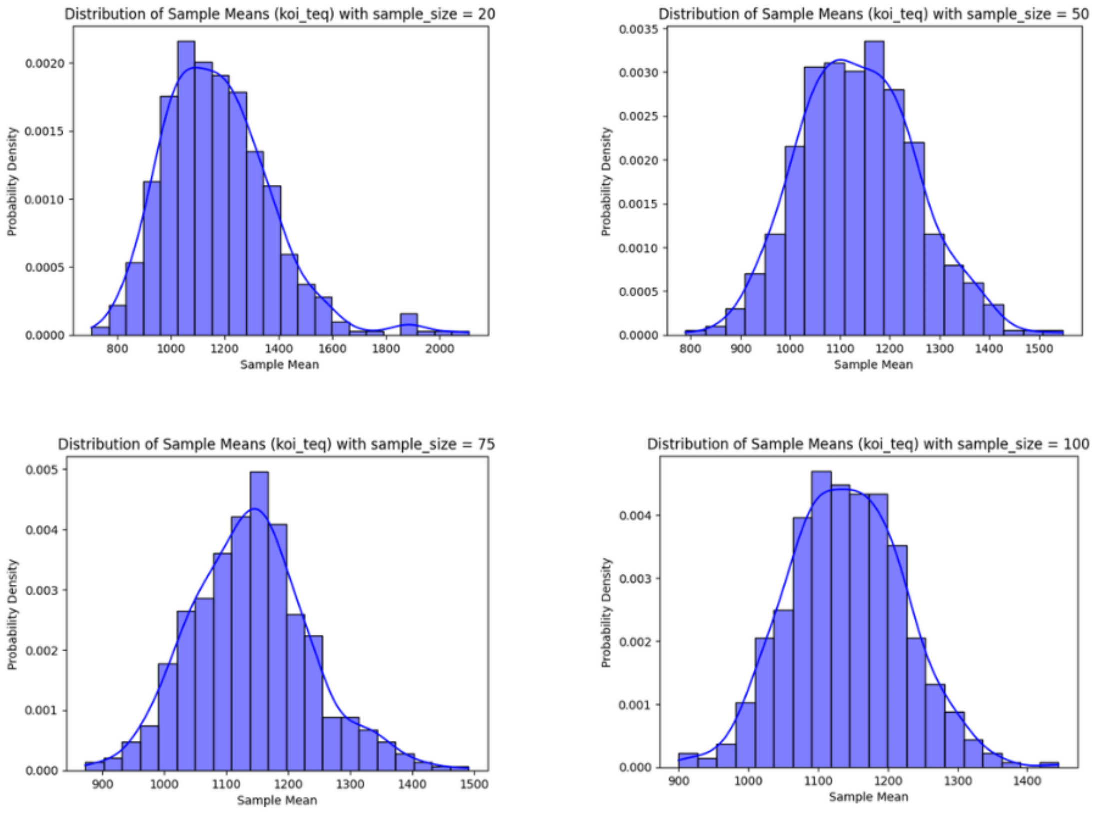

Figure 7: Visualization with varying sample sizes

To gain deeper insights into the characteristics of the various sample sizes, we have compiled the findings into Table 2, presenting their corresponding sample statistics. This comprehensive table allows for a comparative analysis of each sample size against the population mean and standard deviations, as detailed in Table 1.

Population Mean (μ)	**Population Standard Deviation (σ)**
1142.01476	846.24489

Table 8: Population Statistics

Sample Size (n)	**Sample Mean ($\bar{x}$)**	**Sample Deviation (s_x)**	**Standard Error**
20	1142.988	789.59628	174.77
50	1143.26796	821.2933	116.14
75	1150.71336	839.9984	96.99466
100	1142.98782	842.26942	83.2269

Table 9: Sample Statistics

For a sample size of 20, we observe that the sample mean (1142.988) is in close proximity to the population mean (1142.01476). The sample deviation (789.59628) reflects the variability within this smaller sample. The standard error (174.77) highlights the precision of the sample mean, acknowledging the inherent variability when dealing with a limited sample size. With an increase in sample size to 50, the sample mean (1143.26796) continues to align closely with the population mean. The sample deviation (821.2933) exhibits a wider range, capturing more variability. The standard error (116.14) noticeably decreases, indicating enhanced precision in estimating the population mean. The histogram shows a smoother curve, with the distribution further resembling a normal distribution.

Expanding the sample size to 75 results in a sample mean (1150.71336) that converges even more closely to the population mean. The sample deviation (839.9984) indicates variability, and the standard error (96.99466) continues its downward trend, signaling heightened precision. The histogram appears more symmetrical, showcasing a distribution that aligns increasingly with the normal distribution. For a sample size of 100, the sample mean (1142.98782) mirrors the population mean, and the sample deviation (842.26942) captures the variability inherent in a larger sample. The standard error (83.2269) notably decreases, demonstrating a refined precision in estimating the population mean. The histogram displays a smoother curve, approaching the characteristics of a normal distribution.

As we progress from smaller to larger sample sizes, the sample means exhibit improved accuracy in approximating the population mean, and the standard error decreases, indicating enhanced precision. This progression aligns with the fundamental principles of the CLT, reaffirming the reliability of sample means as estimators of population parameters.

3.2 Confidence Interval

Confidence intervals play a pivotal role in the realm of statistical analysis, providing a robust methodology for quantifying the inherent uncertainty within sample estimates and fortifying the credibility of our conclusions. These intervals furnish a spectrum of plausible values for a population parameter, instilling confidence in the accuracy of our statistical estimates. A fundamental principle underlying confidence intervals is the inverse relationship between width and precision—wider intervals correspond to less precise estimates, while narrower intervals signify greater precision. In our exploration, we have constructed diverse confidence intervals to apply to sample data, employing varying confidence levels of 90%, 95%, and 99.5%. This deliberate selection allows us to empirically examine and validate the anticipated relationship between interval width and estimate precision.

Another important metric that allows us to better understand the concept of confidence interval is margin of error. Ity encapsulates the range within which we can reasonably expect the true population mean to lie. Understanding the margin of error allows us to make more informed decisions based on the reliability of data. A larger margin of error suggests a higher level of uncertainty and a need for cautious interpretation of the results.

Sample Size	**Confidence Level**	**Lower Bound**	**Upper Bound**	**Margin of error**
20	90	1129.94	1155.86	12.97
20	95	1127.4531	1158.3846	15.46
20	99.5	1120.72409	1165.1135	22.194
50	90	1134.5569	1151.9879	8.711
50	95	1132.88222	1153.6536	10.38
50	99.5	1128.3463	1158.72	14.9
75	90	1143.496	1158.0466	7.27
75	95	1142.09	1159.4453	8.67
75	99.5	1138.3233	1163.219	12.447
100	90	1136.54865	1149.426	6.439
100	95	1135.31	1150.554	7.677

100	99.5	1131.97	1154.0052	11.01

Table 10: Confidence intervals

As seen in table 3, for the sample size of 20, varying confidence levels yield distinct confidence intervals and associated margins of error. At a 90% confidence level, the interval (1129.94, 1155.86) suggests that we can be 90% confident the true population mean falls within this range. The corresponding margin of error is 12.97, reflecting the potential variability in estimates. As we increase the confidence level to 95% and 99.5%, the intervals (1127.4531, 1158.3846) and (1120.72409, 1165.1135) expand, with larger margins of error (15.46 and 22.194, respectively). This indicates a trade-off: higher confidence demands wider intervals, introducing more uncertainty.

For a sample size of 50, the pattern continues. At a 90% confidence level, the interval (1134.5569, 1151.9879) has a smaller margin of error (8.711) compared to the 95% and 99.5% confidence intervals, demonstrating the inherent balance between precision and confidence. As we progress to sample sizes of 75 and 100, a similar trend emerges. The larger the sample size, the narrower the confidence interval and the smaller the margin of error.

The choice of sample size and confidence level involves a delicate balancing act. Smaller sample sizes yield broader intervals but require less data, while larger sample sizes provide more precise estimates but may be resource-intensive. The selection should align with the research goals and acceptable level of uncertainty. For this particular scenario, a sample size of 100 with a 95% confidence level might strike a reasonable balance between precision and practicality, offering a relatively narrow interval (1135.31, 1150.554) and a manageable margin of error (7.677).

3.3 Hypothesis Testing

Hypothesis testing is a fundamental and powerful statistical tool that allows researchers to draw inferences and make decisions based on data. It involves formulating and evaluating competing hypotheses about population parameters, providing a structured approach to validate assumptions or claims. The process typically begins with a null hypothesis H_0, which represents a default or no-effect assumption, and an alternative hypothesis H_a, positing the existence of a significant effect.

In the context of hypothesis testing, the examination of sample means obtained from different sample sizes (specifically, 20 and 100) as representatives of the population mean will be conducted using a 95% confidence interval. To assess the validity of these sample means, a one-sample t-test will be employed, which compares the means of the samples to the population mean. This test involves formulating a null hypothesis assuming no significant difference between the sample mean and the population mean, and an alternative hypothesis suggesting a significant difference. The test statistic and p-value will be calculated, and the decision to reject the null hypothesis will be made based on the significance level (commonly set to 0.05). If the p-value is less than 0.05, we reject the null hypothesis, indicating that there is sufficient evidence to suggest a significant difference between the sample mean and the population mean. This implies that the observed sample means may not be representative of the entire population of temperature equilibrium, raising questions about the reliability of the sample estimates in this context.

In a two-tailed t-test, the T-statistic is a measure of how far the sample mean deviates from the population mean in terms of the standard error of the mean. The negative or positive sign of the T-statistic indicates the direction of the deviation. The p-value represents the probability of observing a T-statistic as extreme as the one calculated, assuming the null hypothesis is true. In other words, it quantifies the evidence against the null hypothesis. A low p-value, typically below the chosen significance level (e.g., 0.05), indicates that the observed results are unlikely to occur by random chance alone. T-statistic provides information about the size and direction of the difference between the sample and population means, while the p-value helps you determine whether this difference is statistically significant. A lower p-value indicates stronger evidence against the null hypothesis.

3.3.1 Assessing the representativeness of sample size 20

In this hypothesis testing, we want to check if the mean of the sample means for a sample size 20 is a good representative of the population mean. The null and alternative hypotheses can be stated as follows:

H_0: There is no significant difference between the mean of the sample means ($\overline{X}_{20}$) and the population mean (μ)
H_a: There is a significant difference between the mean of the sample means ($\overline{X}_{20}$) and the population mean (μ)

After performing a two-tailed t-test, we get the T-Statistic as -2.379 and p-value of 0.002. A T-Statistic of -2.379 suggests that the sample mean is significantly lower than the population mean. A p-value of 0.002 is lower than the significance level of 0.05. Therefore, we reject the null hypothesis and conclude that the mean of the sample means ($\overline{X}_{20}$) is not a good representative of the population mean of temperature equilibrium.

3.3.2 Assessing the representativeness of sample size 100

In this hypothesis testing, we aim to assess whether the mean of the sample means for a sample size of 100 serves as a good representative of the population mean. The null and alternative hypotheses can be formulated as follows:

H_0: There is no significant difference between the mean of the sample means ($\overline{X}_{100}$)and the population mean (μ).

H_a: There is a significant difference between the mean of the sample means ($\overline{X}_{100}$) and the population mean (μ).

Upon conducting a two-tailed t-test, the obtained T-Statistic is -1.835 and the associated p-value is 0.067. The T-Statistic suggests that the sample mean is slightly lower than the population mean, though not to a statistically significant extent. The p-value of 0.067 is higher than the significance level of 0.05, leading us to fail to reject the null hypothesis. Consequently, we do not find sufficient evidence to conclude that the mean of the sample means ($\overline{X}_{100}$) significantly differs from the population mean of temperature equilibrium.

3.4 Multivariate Analysis

Multivariate analysis is a powerful statistical approach that delves into the intricate relationships and interactions between multiple variables simultaneously. Unlike univariate analysis, which focuses on a single variable, multivariate analysis allows us to explore the complex interplay among several variables within a dataset. This methodology is essential for uncovering patterns, dependencies, and correlations that may not be evident when considering variables in isolation.

In this comprehensive case study, we embark on a journey of multivariate analysis, seeking to unravel nuanced insights from various dimensions of our dataset. Our initial focus centers on a bivariate analysis, specifically comparing two categorical variables: Stellar Eclipse Flag (koi_fpflag_ss) and Centroid Offset Flag (koi_fpflag_co). Through a statistical test, we aim to discern the potential independence or association between these variables. Subsequently, our exploration extends to understanding how the equilibrium temperature (koi_teq) varies across different categories within the Stellar Eclipse Flag (koi_fpflag_ss).

Furthering our investigation, we delve into a comparison of numerical variables—specifically, the Equilibrium Temperature (koi_teq) and the average Stellar Temperature (koi_steff). This analysis involves scrutinizing whether the average values of these variables exhibit statistical significance from one another. To encapsulate the complexity of relationships within our dataset, we then venture into two distinct regression analyses: a simple linear regression and a multi-linear regression. These regression models aim to elucidate the intricate interplay and predictive patterns among variables, providing a comprehensive understanding of the dataset's dynamics. Through these diverse multivariate analyses, we aspire to glean meaningful insights and contribute to a holistic understanding of the dataset under scrutiny.

3.4.1 Stellar Eclipse Flag vs Centroid Offset Flag

Stellar Eclipse Flag (koi_fpflag_ss) and Centroid Offset Flag (koi_fpflag_co) are two categorical variables that play a crucial role in characterizing the observed exoplanetary candidates. The Stellar Eclipse Flag flags potential issues related to stellar eclipses, indicating whether there is a potential contamination or confusion due to the presence of a nearby star. On the other hand, the Centroid Offset

Flag signifies the presence of a centroid offset, hinting at a potential discrepancy in the positioning of the observed transit signal compared to the centroid of the target star.

In speculating their potential relationship, it can be inferred that both flags might be intertwined in scenarios where a stellar eclipse and centroid offset occur simultaneously. For instance, if a flagged stellar eclipse leads to a positional shift in the centroid, it could trigger the activation of both flags. Conversely, their independence could also be plausible, suggesting that a flagged stellar eclipse may not necessarily coincide with a centroid offset and vice versa. By comparing these two flags, we aim to uncover patterns or disparities that could enhance our understanding of the observational data and contribute to refining the criteria for identifying potential exoplanetary candidates.

To assess the relationship between the two categorical variables, Stellar Eclipse Flag and Centroid Offset Flag , we opted for the Chi-Square test of Independence. A crucial prerequisite for this test is that all expected frequencies, obtained by cross-tabulating the categorical variables against each category, must surpass a threshold of 1. After performing the cross-tabulation, as illustrated in Table 4, it is evident that all the expected frequencies exceed the minimum value of 1. This adherence to the assumption ensures the validity of our Chi-Square test, allowing us to proceed with confidence in examining the independence or potential association between Stellar Eclipse Flag and Centroid Offset Flag.

Expected Frequencies	**koi_fpflag_co = 0**	**koi_fpflag_co = 1**
koi_fpflag_ss = 0	4553.21	1295.71
koi_fpflag_ss = 1	1669.79	475.21

Table 11: Expected frequencies of koi_fpflag_ss v koi_fpflag_co

The Chi-Square test of Independence conducted on the categorical variables Stellar Eclipse Flag and Centroid Offset Flag yielded a substantial Chi-Square Statistic of 130.99 with an extremely low p-value (2.49e-30). This p-value falls well below the conventional significance level of 0.05, leading us to reject the null hypothesis. The rejection of the null hypothesis implies a significant association between the two categorical variables, indicating that the occurrence of one flag is not independent of the other. In statistical terms, the Chi-Square Statistic quantifies the extent of the observed association, and the exceedingly low p-value supports the assertion that this association is unlikely to be a result of random chance. Therefore, these results provide valuable insights to a statistician, suggesting that changes in one flag category are indicative of changes in the other, and vice versa, offering a basis for further investigation into the underlying factors contributing to this observed dependency.

The initial speculation regarding the potential interdependence of Stellar Eclipse Flag and Centroid Offset Flag has been substantiated through rigorous statistical analysis. The flags appear to share a significant relationship, suggesting that occurrences of one flag may be closely associated with the presence of the other.

3.4.2 Equilibrium Temperature vs Stellar Eclipse Flag

In our case study to explore the potential variations in equilibrium temperature corresponding to different categories of the Stellar Eclipse Flag (koi_fpflag_ss), we embark on a statistical examination. Specifically, we aim to investigate whether the occurrence of equilibrium temperature is contingent upon the particular category of the stellar eclipse flag. To evaluate this hypothesis, we employ the Kruskal-Wallis Test, a robust non-parametric method. The choice of this test is motivated by the observation that the distribution of equilibrium temperature does not conform to the normal distribution, necessitating a robust approach to discern potential differences across the categories of the stellar eclipse flag. Through the application of the Kruskal-Wallis Test, we seek to uncover any statistically significant variations in equilibrium temperature associated with distinct categories of the Stellar Eclipse Flag. By exploring how the equilibrium temperature varies with the presence or absence of a stellar eclipse, astronomers can gain insights into the thermal dynamics of exoplanetary systems.

Prior to conducting the Kruskal-Wallis test, we categorized the data into two groups based on the Stellar Eclipse Flag values: one group comprising temperature equilibrium values where koi_fpflag_ss = 1, and the other with values where koi_fpflag_ss = 0. Subsequently, the Kruskal-Wallis test was employed to assess whether there exists a significant difference in the equilibrium temperatures between these groups. The obtained statistic of 328.18875 and a remarkably low p-value of 2.38e-73, well below the chosen significance level, signify a substantial difference in the temperature distributions across the two groups. This suggests that the occurrence of a stellar eclipse, as indicated by the Stellar Eclipse Flag, is associated with a notable variance in equilibrium temperatures. The Kruskal-Wallis test, suited for non-normally distributed data, reaffirms the statistical significance of this observation. In conclusion, we reject the null hypothesis, confirming a significant difference in equilibrium temperatures based on the presence or absence of a stellar eclipse, providing valuable insights into the thermal dynamics of exoplanetary systems.

3.4.3 Equilibrium Temperature vs Stellar Effective Temperature

Comparing the Stellar Effective Temperature (koi_steff) and Equilibrium Temperature (koi_teq) holds significant relevance in understanding the relation between the properties of host stars and the equilibrium conditions of orbiting exoplanets. The Stellar Effective Temperature serves as a crucial indicator of the host star's intrinsic luminosity and energy output, influencing the thermal environment of orbiting planets. On the other hand, the Equilibrium Temperature reflects the balance between the incoming stellar radiation and the thermal emission from the exoplanet. By scrutinizing the relationship between these two numerical variables, we aim to unveil potential correlations that shed light on the thermal regulation mechanisms within exoplanetary systems.

To assess the potential difference in means between Stellar Effective Temperature and Equilibrium Temperature , we opted for the non-parametric Wilcoxon Signed-rank test. This choice aligns with the acknowledgment that these variables do not conform to a normal distribution. Upon conducting the test, the obtained statistic is 70500, and the associated p-value is 0. Given that the p-value falls below the chosen significance level, we reject the null hypothesis, concluding that there is a significant difference between the means of equilibrium temperature and stellar effective temperature. The Wilcoxon Signed-rank test, which assesses the distribution of differences between paired observations, affirms the existence of a notable distinction in this case.

3.4.4 Regression Analysis

3.4.4.1 Simple Linear Regression

Simple linear regression is a statistical method that allows you to summarize and study the relationship between two continuous (quantitative) variables. It assumes that there is a linear relationship between the independent variable (predictor) and the dependent variable (response). The goal of simple linear regression is to fit a linear equation to the observed data that best describes the relationship between the variables.

The linear equation has the form: $Y + \beta_0 + \beta_1 X + \varepsilon$

Where :

Y is the dependent variable (response)

X is the independent variable (predictor),

β_0 is the intercept,

β_1 is the slope, representing the change in Y for a one-unit change in X,

ε is the error term, accounting for unobserved factors affecting

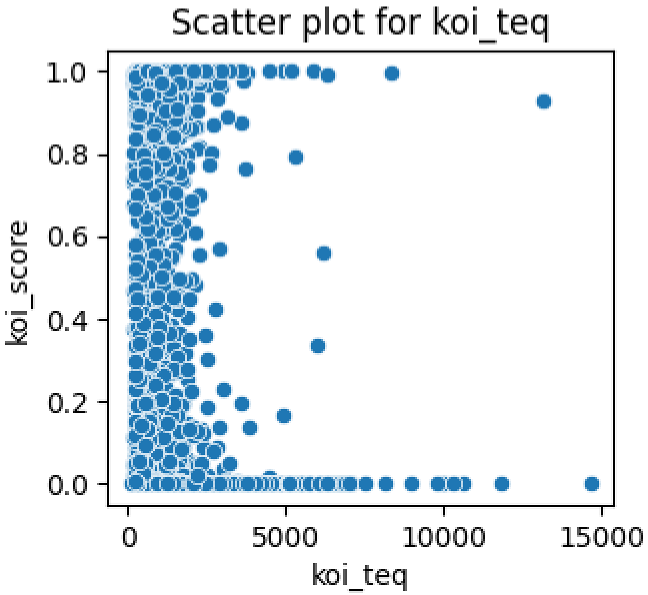

Figure 8: Scatter plot showing relationship between koi_teq and koi_score

For our analysis we consider Equilibrium Temperature (koi_teq) vs Disposition Score (koi_score). From the Scatter plot we see that there is little to no correlation between Disposition Score and Equilibrium Temperature. Most koi_score values are concentrated at the lower range of koi_teq, and there's a large variance in koi_score for low koi_teq values.

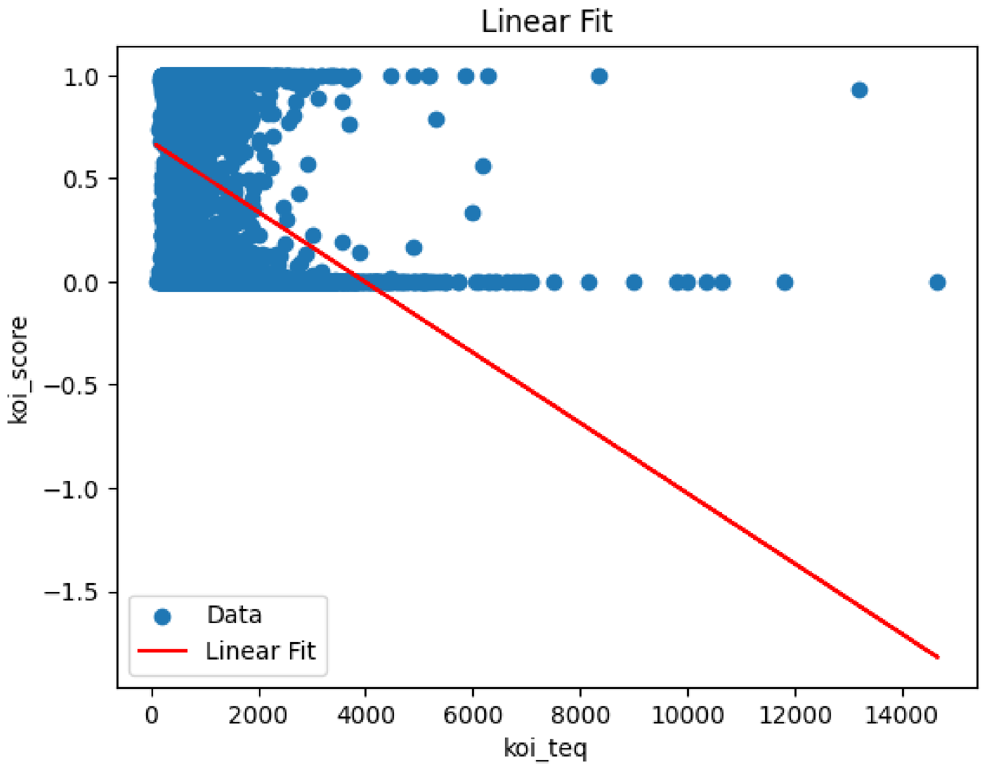

Figure 9: Fitted Scatter plot

From the linear fit, we found that there is a negative correlation between koi_score and koi_teq. As 'koi_teq' increases, the 'koi_score' tends to decrease. The slope of the line is negative, -0.000173710094358967, which quantifies the decrease in koi_score for each unit increase in koi_teq. The intercept value is approximately 0.678, suggesting that when koi_teq is zero, the koi_score would be around 0.678.

We performed Ordinary Least Squares (OLS) Regression method and achieved following results:

R^2	F-statistic	F-stats P value	Coeff X_1	T-statistic	T-stats P value	Mean Squared Error
0.044	4553.21	2.14e-64	-0.1420	-17.136	0.00	0.44725443937

Table 12: Ordinary Least Squares Regression results

The model has a very low R-squared value of 0.044, which indicates that the independent variable explains only 4.4% of the variance in the dependent variable, suggesting a weak fit. The F-statistic is significant (with a very low p-value), indicating that the overall regression is statistically significant. The coefficient for X_1 is -0.1420, and it is statistically significant given the t-statistic of -17.136 and a p-value of 0.000. The Mean Squared Error (MSE) is 0.447254439375. This indicates the average squared difference between the observed actual outcoming values and the values predicted by the model. The lower the MSE, the better the model is at predicting outcomes.

In conclusion, while the model is statistically significant, its predictive power is weak, and there are indications of potential issues with the residuals.

3.4.4.1.1 Assumptions made

We made the following assumptions with regards to our selected data -
Linearity: The relationship between the independent variables and the dependent variable is linear. This assumption is violated here, as the scatter plot does not indicate a linear relationship.

Independence: The residuals (errors) are independent.

Homoscedasticity: The residuals have constant variance at every level of the independent variable. The scatter plot suggests potential heteroscedasticity since the variance of koi_score seems to increase with koi_teq.

Normality of Residuals: The residuals are normally distributed. The Jarque-Bera test in the first image suggests this assumption is violated.

No significant outliers, **high leverage points**, or **influential data points**: These can strongly impact the regression line and violate OLS assumptions.

3.4.4.2 Multiple Linear Regression

Multiple linear regression is an extension of simple linear regression, where the goal is to model the relationship between a dependent variable (target) and two or more independent variables (features or predictors). In other words, it allows you to analyze the impact of multiple variables on the outcome.

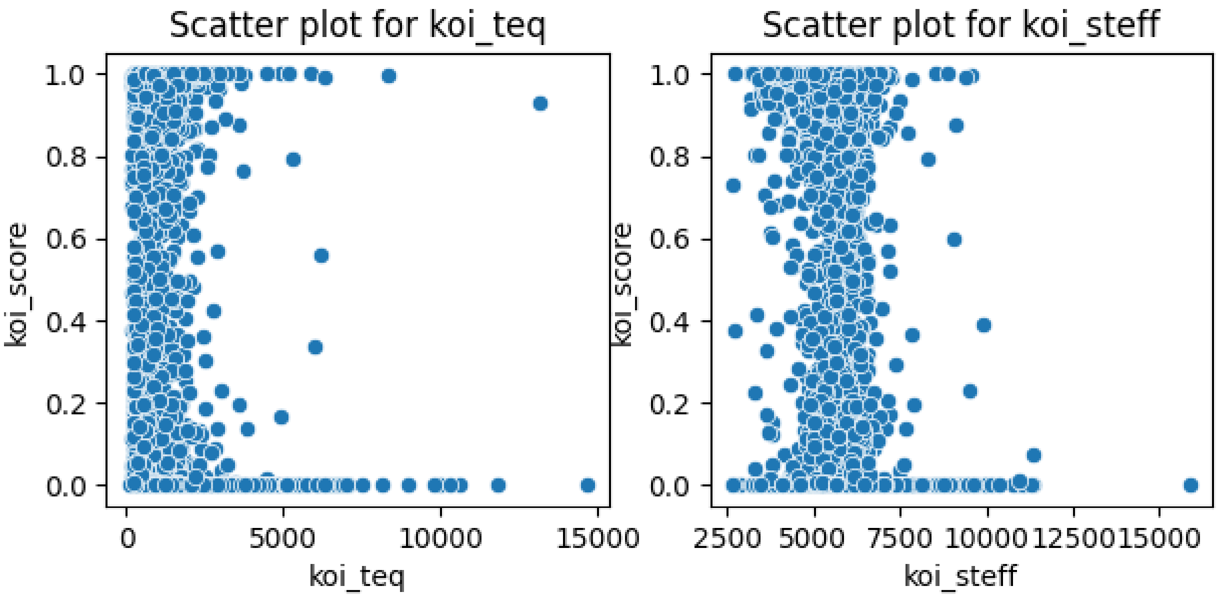

Figure 10: Scatter plot showing relationship between koi_score vs koi_teq and koi_score vs koi_steff

We then selected two independent variables (koi_teq and koi_steff), and the dependent variable (koi_score). This is what we could infer from the above plots - Scatter Plot for koi_teq shows that the points are more spread out as koi_teq increases, showing a wide variation in koi_score at higher koi_teq values. There doesn't appear to be a clear trend or pattern indicating a strong linear relationship between koi_teq and koi_score. Scatter Plot for koi_steff shows that the data points are densely packed towards the lower end of koi_steff, indicating a large number of observations with lower koi_steff values.Similar to the first plot, there isn't a clear linear trend visible, which suggests that koi_steff' may not have a simple linear relationship with koi_score.

We performed Ordinary Least Squares (OLS) Regression method and achieved following results:

	R^2	F-statistic	F-stats P value	Coeff	T-statistic	T-stats P value	Mean Squared Error
x_1	0.051	171.4	2.99e-73	-0.1271	-14.895	0.00	0.44249977315
x_2	0.051	171.4	2.99e-73	-0.0586	-6.862	0.00	0.44249977315

Table 13: Ordinary Least Squares Regression results

We can infer the following from the OLS regression results - Both predictors (koi_teq and koi_steff) are negatively associated with the dependent variable, with x1 having a coefficient of -0.1271 and x2 -0.0586. P-Values: The coefficients for both x1 and x2 are statistically significant ($p < 0.001$), suggesting a strong evidence against the null hypothesis (which typically states there is no effect). The R-squared value is 0.051, which means that approximately 5.1% of the variability in koi_score is explained by the model. This is generally considered a low value, indicating that the model does not explain much of the variation in the dependent variable. The F-statistic is significant ($p < 0.001$), indicating that the model is a better fit than an intercept-only model.

In summary, while the model shows significant predictors, its explanatory power is quite low (as evidenced by the low R-squared value), and there are indications that the residuals may not be normally distributed, which could be a violation of one of the assumptions of OLS regression.

3.4.4.3 Simple vs Multiple Linear regression.

1. Coefficients:

In simple linear regression, the coefficient for equilibrium temperature is significant (based on t and p-values), suggesting a strong relationship with the dependent variable.

In multiple linear regression, the coefficients for both equilibrium temperature and stellar flag are significant, indicating that both have an influential relationship with the dependent variable.

2. Model Fit:

The R-squared value, which represents the proportion of variance for the dependent variable explained by the independent variables, is higher in the multiple regression model than in the simple regression. This suggests that the multiple regression model explains the variability of the output variable better than the simple regression model.

3. F-statistic:

The F-statistic in both models indicates that the models are statistically significant overall.

From these results, we can infer that adding the stellar flag as a second independent variable in our regression model improves the model's explanatory power for the dependent variable. The significant coefficients for both predictors in the multiple regression model suggest that both equilibrium temperature and stellar flag have a notable impact on the dependent variable, with the multiple regression model providing a more comprehensive understanding of the relationship between the variables.

4 Conclusion

In concluding this comprehensive analysis of the NASA Exoplanet Archive dataset, we have employed a variety of statistical techniques to explore the intricacies of exoplanetary data. The univariate analysis reinforced the Central Limit Theorem and provided confidence in sample estimations, while hypothesis tests offered insightful validations. The multivariate analysis, including both simple and multiple linear regressions, unearthed significant relationships between equilibrium temperature, stellar flag, and the dependent variable. Notably, the inclusion of the stellar flag in the multiple regression model improved its explanatory power, albeit the overall predictiveness remained moderate as indicated by low R-squared values.

Furthermore, the Chi-Square tests revealed a significant association between stellar eclipse and centroid offset flags, suggesting interconnectedness in observational data. The Kruskal-Wallis and Wilcoxon Signed-rank tests identified significant differences across groups, highlighting the influence of stellar phenomena on equilibrium temperatures and corroborating the thermal dynamics within exoplanetary systems.

Despite achieving statistically significant results, several limitations were encountered. Assumptions of normality and homoscedasticity were challenged, and low R-squared values pointed to a limited capacity of our models to account for the variability in the dependent variable. These insights emphasize the complex and multifactorial nature of exoplanetary environments and the necessity for advanced modeling techniques to capture their dynamics more effectively.

In summary, this analysis has elucidated several key factors that contribute to the characterization of exoplanets, shedding light on the complex relationships between these celestial bodies and their host stars. The findings provide a foundation for future research, suggesting that additional variables and possibly non-linear models may yield improved understanding and prediction of exoplanetary attributes.