Task 2.1 Hypothesis formulation

Dependent and Independent Variables:

In mathematics and statistics, the terms "dependent variable" and "independent variable" are used to describe the relationship between two variables.

An independent variable is a variable that is manipulated or controlled by the researcher. It is also called the "input variable" as it is the input to the system. The value of the independent variable does not depend on any other variable in the system.

On the other hand, a dependent variable is a variable that is being studied and measured. It is also called the "output variable" as it is the output of the system. The value of the dependent variable depends on the value of the independent variable. In other words, the dependent variable changes in response to changes in the independent variable.

Here as we have data from 5000 companies and it has 19 templates as shown in the ipynb file.As many of these columns are number, id, metro, source, city, page URL and URL etc which we don't need because they don't depend on each other and we cannot even compare them. While we talk about dependent and independent variables columns like Growth Revenue and workers are highly dependent on each other and they can have a positive and negative relationship with each other which we show in further in the data using code. If we talk about our dependent variable here, it can be revenue because the increase and non-increase in workers and growth affect revenue. Since we are treating growth and workers as independent variables.

Now we compare the three variables workers' growth and revenue using the correlation function in Python and also visualize these variables with the help of a heatmap graph. As we compared the three metric columns in the notebook. Resulting in a positive relationship between the workers with the revenue and worker has a negative relationship with growth. While growth and revenue have a positive relationship and their reflection is also shown in regularly. We have tested this hypothesis at the level of relationship. As a result, we can assume that if the number of workers in our company is high, our company's growth will be negative and it will have a negative impact on our company's growth as we have seen in correlation Which we find out with the help of the correlation function of python and we can also show it with heatmap graph. We can also estimate that if

the number of workers in our company is more, the labour rate of the company will also increase, due to which the company will not be able to and the growth rate of the company will be low. But if the number of workers is more then the revenue of the company will increase because people will be working in the company so their sales will be more but their profit will be negatively affected.

When we compare the revenue column, it can also be estimated which is also proved by the correlation that we have a positive relationship with growth and workers which means that if a company has revenue increases then the number of its workers will also increase and that company will also be doing every year. Therefore Revenue has a positive relationship with workers and growth rate and it is also understood in our common sense.

We also have state-wise data of the united states as categorical data which includes all US states including California, Georgia, New York and Alaska etc. We have only extracted data from California states, which shows that companies here have revenue of more than 52M While a few companies have a rating of two or six. But there is an interesting thing about it which is also related to our hypothesis that Revenue had a positive relationship with the workers and the growth rate and the relationship between growth rate and workers were negative. If we look carefully at this table, we know that The number of workers in companies in California is in the hundreds and the revenue also goes from 50 million dollars which proves that our hypothesis is valid. Because revenue has a positive relationship with growth and workers and a negative relationship between growth and workers.

The purpose of taking growth revenue and workers in correlation was that they were directly influencing each other positively and negatively and these three columns are enough for our comparison while the rest of about ten numeric columns because we have no need as such because these columns are never influencing each other directly or indirectly.

Note:

We have done the data cleaning work earlier due to which there are showing only three target columns showing and you can clearly check this in the notebook. You can clearly check the correlation and heatmap graph by going to the notebook with the task number.

2.2 Data Preparation

Here we have to discuss about data Preparation we have data about companies in the data there are 5000 rows and 19 columns first we identify those columns which has no effect or importance on the data. They include input, widget, id, source, page URL, metro, city and rank etc. We removed them from the data. Now we have only 6 columns. There are Revenue, growth, workers, state, company and industry-related columns include. All this we are doing data cleaning work. There are no null values in the data, meaning zero percent of null values in the data which we checked with Python. After that, we tried to find out the highest number of workers that a company has. For this, we used the python library pandas max function to find the maximum number of workers. From which we know that a maximum of more than 34000 workers work in a company. Duplicate values in the data were also checked And duplicate values in our data were also zero percent. We then checked our data to see if it was normal or not. So for this, we applied the skew function to the data This shows that our data is positively skewed. We have made it clear below by creating a graph heat map which also shows data is positively skewed. In all these steps we checked and tried to clean the data.

Now the data transformation process has started in which we have two columns revenue and growth involved. We want to transform these two columns. We want to bring the revenue column to the million digits. while We want to bring the revenue column to the million digits So that it is easy to understand. We simply divided one million into the revenue column and created a new column by applying the round function. Similarly, taking the growth total to convert the growth column to a percentage divided by each company's growth rate. This gives us another new column for growth Percentage. Now we have also transformed our data. We have again dropped the irrelevant columns once again. Now we have revenue and growth in percentage and million .which came to us as a new column so we deleted the old Growth and Revenue column. We still have six columns as before and it includes revenue and growth.

As mentioned above we have positively skewed data and can't even transform such as. Due to this, we have to apply a non-parametric statistical test to them. In which is the first test is the Mann-Whitney test that we applied with the help of a Python library scipy. We have also unpaired data. In this, we compared the two columns growth and revenue which also gave us a p-value of zero. Which means we have to reject null hypothesis. We applied another group Kruskal wills H test on it. In this also the same two columns revenue and group were tested.In this one, we also got a p-value zero, which means we used the correct tests

Data Transformation:

Data transformation refers to the process of converting data from one format or structure to another, with the aim of making it more suitable for analysis, modeling, or other downstream tasks. Here are some of the benefits of data transformation:

Benefits of Data Transformation

- **Data transformation can help improve the quality of data by eliminating inconsistencies, redundancies, and errors in the data.**
- **When data is collected from different sources, it may be in different formats or structures. Data transformation can help integrate data from different sources by converting it into a common format or structure.**
- **More efficient data analysis: Data transformation can help make data analysis more efficient by reducing the time and effort required to extract insights from the data.**
- **Data transformation can help improve the accuracy of data by removing outliers and other anomalies that can distort the results of data analysis.**
- **Data transformation can help decision-makers make better decisions by providing them with more accurate and relevant information.**
- **Data transformation can help make data more visually appealing and easier to understand by converting it into a format that can be easily visualized, such as a graph or chart.**

Overall, data transformation plays a critical role in the data analytics process, helping organizations make better use of their data and derive more meaningful insights from it.

Task 2.3 Data analysis

The Mann-Whitney test and Kruskal-Wallis H test are both non-parametric tests used to compare two or more independent groups. Here are some general guidelines to help determine when to use each test:

Mann-Whitney test:

- **Used to compare the median of two independent groups.**
- **Assumes that the data are independent and non-normally distributed.**
- **Suitable for analyzing ordinal or continuous data.**

Kruskal-Wallis H test:

- **Used to compare the medians of three or more independent groups.**
- **Assumes that the data are independent and non-normally distributed.**
- **Suitable for analyzing ordinal or continuous data.**
- **The Kruskal-Wallis H test is an extension of the Mann-Whitney test, which is used to compare only two groups.**

Here below are several reasons why we should choose to use these tests instead of parametric tests such as the t-test or ANOVA:

1. **The Mann-Whitney test and Kruskal-Wallis H test are robust to non-normal distributions, which is often the case in real-world data. Unlike parametric tests, these tests do not require the assumption of normality, making them more suitable for skewed or heavy-tailed data.**
2. **The Mann-Whitney test and Kruskal-Wallis H test are suitable for analyzing ordinal data. Parametric tests, such as the t-test or ANOVA, require continuous data and assume that the differences between the groups are normally distributed. In contrast, the Mann-Whitney test and Kruskal-Wallis H test can be used to analyze data that are measured on an ordinal scale, where the differences between values may not be uniform or equally spaced.**
3. **The Mann-Whitney test and Kruskal-Wallis H test are appropriate when the data are independent, and the assumption of independence is less restrictive than the assumption of normality. These tests can be used to analyze data from randomized controlled trials, observational studies, or other experimental designs where the groups are independent.**
4. **The Mann-Whitney test and Kruskal-Wallis H test are less sensitive to outliers than parametric tests, making them more robust to extreme values that may be present in the data.**

The Mann-Whitney test and Kruskal-Wallis H test are both useful statistical tests for analyzing a wide range of data, including our company data. Here are some reasons of how these tests applied to our company dataset:

1. **Salary comparison: A company may want to compare the salaries of two groups of employees, such as male and female employees, to determine if there is a significant difference in pay. The Mann-Whitney test can be used to compare the median salaries of the two groups, and determine if the difference is statistically significant.**

2. **Employee performance: A company may want to compare the performance of three or more teams or departments, to determine if there is a significant difference in their performance. The Kruskal-Wallis H test can be used to compare the median performance scores of each group and determine if there is a significant difference between the groups.**
3. **Customer satisfaction: A company may want to compare the satisfaction levels of two or more customer groups, such as new and existing customers, to determine if there is a significant difference in their satisfaction levels. The Mann-Whitney test or Kruskal-Wallis H test can be used to compare the median satisfaction scores of each group and determine if there is a significant difference.**
4. **Product quality: A company may want to compare the quality of two or more products, to determine if there is a significant difference in their quality ratings. The Mann-Whitney test or Kruskal-Wallis H test can be used to compare the median quality scores of each product and determine if there is a significant difference.**

In all these examples, the Mann-Whitney test and Kruskal-Wallis H test are useful for comparing groups when the data is non-normally distributed, or when the groups are measured on an ordinal scale. These tests provide a robust and reliable way to analyze data and draw meaningful conclusions about the differences between groups.

The Kruskal-Wallis H test and Mann-Whitney test are both non-parametric tests used to compare two or more independent groups when the data is not normally distributed or the assumptions of parametric tests are violated.

The Kruskal-Wallis H test is used to determine whether there is a significant difference between three or more independent groups. It ranks the values from all groups, calculates the sum of ranks for each group, and uses these values to determine if there is a significant difference between groups. The null hypothesis is that the medians of all groups are equal, while the alternative hypothesis is that at least one group differs significantly from the others.

On the other hand, the Mann-Whitney test is used to compare two independent groups. It ranks the values from both groups, calculates the sum of ranks for each group, and uses these values to determine if there is a significant difference between groups. The null hypothesis is that the medians of both groups are equal, while the alternative hypothesis is that one group differs significantly from the other.

Both tests provide a p-value that indicates the probability of observing the test statistic under the null hypothesis. If the p-value is less than the level of significance (usually 0.05), we reject the null hypothesis and conclude that there is a significant difference between groups.

The implications of the test outcomes for answering the business question depend on the specific research question and the context of the study. For example, if a business wants to compare the sales of three different products in different regions, the Kruskal-Wallis H test can be used to determine if there is a significant difference between the sales of the products. If the test is significant, further post-hoc tests such as the Dunn's test can be conducted to identify which groups differ significantly from the others.

Similarly, the Mann-Whitney test can be used to compare the performance of two different marketing campaigns or to compare the salaries of male and female employees in a company. If the test is significant, it implies that there is a significant difference between the groups being compared.

In summary, both Kruskal-Wallis H test and Mann-Whitney test are powerful non-parametric tests that can be used to compare independent groups. However, their implications for answering a business question depend on the specific research question, the context of the study, and the post-hoc analyses conducted after the tests.

There are several ways to show a graphical presentation of the Mann-Whitney test and Kruskal-Wallis H test outcomes, as well as to interpret them. Here are some options:

1. **Box plot: A box plot is a common way to display the distribution of data and compare medians between groups. The box plot displays the median, the interquartile range (IQR), and the range of data for each group. The whiskers extend to the highest and lowest observations that are within 1.5 times the IQR from the box. Any observations outside the whiskers are considered outliers. This plot can be used for both Mann-Whitney test and Kruskal-Wallis H test.**
2. **Scatter plot: A scatter plot can be used to visualize the distribution of data points for both groups in the Mann-Whitney test. Each point represents an observation and the position on the x-axis indicates the value of the variable for group 1, while the position on the y-axis indicates the value of the variable for group 2. A horizontal line can be added to the plot to indicate the median value for each group.**
3. **Bar plot: A bar plot can be used to visualize the median and interquartile range for each group in the Kruskal-Wallis H test. This plot displays the median as a horizontal line within a box that extends from the 25th to 75th percentiles of the data for each group.**

4. **Pair Plot:A pair plot, also known as a scatter plot matrix, is a graphical tool used to explore the relationship between multiple variables in a dataset. Pair plots are particularly useful when dealing with high-dimensional datasets, as they allow the visualization of the pairwise relationships between variables. Pair plots can help identify potential patterns or trends in the data, such as correlations, clusters or outliers. They are commonly used in exploratory data analysis to gain insights into the data, and can also be used to visualize the results of a statistical analysis.**

One of the limitations of the Kruskal-Wallis H test is that it assumes that the samples being compared have the same shape of distribution. If this assumption is violated, the test may not be accurate. Additionally, the Kruskal-Wallis H test assumes that the samples being compared have the same variance, which may not be the case in some situations.

Similarly, the Mann-Whitney test assumes that the samples being compared have the same shape of distribution, and that the samples are independent and randomly selected. Violations of these assumptions can lead to inaccurate results.

To enhance the accuracy of the Kruskal-Wallis H test and Mann-Whitney test, researchers can consider using methods such as bootstrapping or permutation testing. Bootstrapping involves resampling the data multiple times to create new samples, which can help to estimate the distribution of the test statistic and the confidence intervals. Permutation testing involves randomly permuting the data and recalculating the test statistic multiple times to estimate the null distribution. Both of these methods can be used to improve the accuracy of the test and to account for violations of the assumptions.

Another way to enhance the accuracy of these tests is to use more advanced statistical methods, such as generalized linear models or mixed effects models. These methods can account for non-normality, unequal variances, and other violations of assumptions, and can provide more robust and accurate results.

Overall, while the Kruskal-Wallis H test and Mann-Whitney test have some limitations and assumptions, they are still useful tools for analyzing non-parametric data. By acknowledging these limitations and using appropriate methods to enhance accuracy, researchers can obtain more reliable and informative results.

Task 2.4 Deployment considerations:

We do extract insights from data by applying statistical and computational methods. Deploying data analytics within a company's dataset involves several steps:

1. **Data collection: Collecting relevant data from publically available datasets within such as databases, spreadsheets, and other data repositories.**
2. **Data cleaning: Cleaning the collected data to remove errors, inconsistencies, and duplicates, and ensuring that the data is accurate and complete.**
3. **Data preparation: Preparing the data for analysis by transforming it into a format that can be easily analyzed by data analytics tools. This may involve data normalization, data aggregation, and data reduction.**
4. **Data analysis: Analyzing the data using various statistical and computational techniques to uncover insights and patterns. This may involve data visualization, clustering, regression analysis, and machine learning.**

5. **Insights and recommendations: Presenting the insights and recommendations to decision-makers within the company, such as executives, managers, and analysts.**
6. **Implementation: Implementing the insights and recommendations within the company to improve operations, reduce costs, or increase revenue.**

The process involves identifying the relevant data sources, cleaning and preparing the data, analyzing the data using appropriate techniques, and presenting the insights and recommendations to decision-makers.

The analysis of risks and potential challenges associated with deployment are done using the following steps:

1. **The first step is to identify all potential risks and challenges that may arise during the deployment process. This is done by reviewing previous deployments, consulting with subject matter experts, and brainstorming with the team.**
2. **First all potential risks and challenges have been identified, and then we next take the step to assess the potential impact of each one. This will be done by evaluating the likelihood of the risk occurring and the severity of its impact.**
3. **After assessing the impact of each risk and challenge, it's important to prioritize them based on their potential impact on the deployment. This will help the team focus on addressing the most critical risks and challenges first.**
4. **Once the risks and challenges have been prioritized, the team can begin to develop risk mitigation strategies. This may involve developing contingency plans, implementing additional controls, or modifying the deployment plan to minimize the potential impact of the risk.**

5. **Monitor and review risks and challenges: The final step is to continually monitor and review the risks and challenges throughout the deployment process. This will help the team to identify any new risks or challenges that may arise and adjust their risk mitigation strategies accordingly**

The key to effectively analyzing risks and potential challenges associated with any deployment is to take a proactive approach and to continuously monitor and review the risks and challenges throughout the process.

The conclusions and recommendations for deployment on this company dataset project and the analysis of the risks and potential challenges associated with it. However, some general recommendations that can be applied to most deployment projects are:

1. **The deployment of a project involves collecting and processing personal data, which can raise ethical concerns related to privacy and data protection. The project team must ensure that appropriate measures are in place to protect individuals' privacy and data, such as obtaining consent, implementing data security measures, and anonymizing data where possible.**
2. **Project deployment should not lead to the creation or perpetuation of bias and discrimination against certain groups of people. This includes ensuring that the project does not reinforce existing biases or stereotypes, and that it is inclusive and fair to all stakeholders.**
3. **The project team should be transparent about the purpose and scope of the project, as well as any risks or potential negative impacts. They should also be accountable for the project's outcomes and take responsibility for any negative consequences that may arise.**

4. **The deployment of a project should promote fairness and equity by ensuring that benefits and burdens are distributed fairly among stakeholders. This includes ensuring that the project does not disproportionately benefit certain groups or individuals over others.**
5. **The project team should consider the environmental impact of the project and take steps to minimize its ecological footprint.**
6. **The deployment of a project may involve the creation or use of intellectual property, such as patents, copyrights, or trademarks. The project team should ensure that they respect the intellectual property rights of others and that they do not infringe on any existing patents or copyrights.**

The ethical aspects of the deployment of a project should be carefully considered and addressed to ensure that the project is deployed in a responsible and ethical manner.

The conclusions and recommendations for deployment should be given below:

1. **A detailed deployment plan that outlines the steps and timelines for the deployment should be developed. The plan should include risk mitigation strategies and contingency plans to address potential challenges that may arise during the deployment.**
2. **The deployment plan should be thoroughly tested to ensure that it is feasible and effective. This may involve conducting a pilot deployment to identify and address any issues before the full deployment.**

3. **Effective communication with stakeholders is essential to the success of the deployment. All stakeholders should be informed about the purpose, scope, and timeline of the deployment, as well as any potential risks or challenges.**
4. **The deployment should be closely monitored and evaluated to ensure that it is progressing as planned and to identify any issues or challenges that arise. This will allow the project team to make adjustments and modifications as needed to ensure a successful deployment.**
5. **Ethical considerations, such as privacy, bias, and environmental impact, should be carefully considered and addressed during the deployment process.**

In conclusion, a successful deployment requires careful planning, effective communication, and ongoing monitoring and evaluation. By following these recommendations, the project team can help to ensure that the deployment is successful and that any potential risks or challenges are addressed in a responsible and ethical manner.

A well-defined model for deployment should help ensure that the deployment process is consistent across different projects, teams, and locations. This will help to reduce errors, improve efficiency, and increase the quality of the deployments.This will also help to minimize the impact of any issues that arise during deployment, reducing costs and minimizing disruption.

—----------