Akshaykumar Vijay Thakare

Student id: 200523970

**Dataset Exploration Project Part 4 (DE pt4) – Inferential Techniques & Final Report**

**Global YouTube Statistics Dataset**

This dataset which is obtained from Kaggle shows detailed statistics of YouTube. It has many key features which provide useful information for data analysis purpose. Overall, it has subscribed YouTube channels along with their name, rank and country also how many uploads and views it has. To add more it also consists earning of each channel.

Through this dataset many potential analyses on several topics can be obtained. Not only analysis but visualization can also be done. This data set has 996 rows and 16 columns.

|  |  |  |
| --- | --- | --- |
| **Field** | **Field Description** | **Data Type** |
| rank | The world rank of YouTube channel | int |
| Youtuber | The name of YouTube channel | char |
| subscribers | Number of subscribers of a channel | int |
| video views | Number global views on the channel | char |
| category | Category of a channel | char |
| uploads | Number of uploads of a channel | int |
| Country | Origin country of a channel | char |
| Abbreviation | Abbreviation of a channel | char |
| country\_rank | Rank in its origin country | int |
| lowest\_monthly\_earnings | Predicted lowest monthly earnings | int |
| highest\_monthly\_earnings | Predicted highest monthly earnings | int |
| lowest\_yearly\_earnings | Predicted lowest yearly earnings | int |
| highest\_yearly\_earnings | Predicted highest yearly earnings | int |
| created\_year | The year when the channel was created | date |
| Population | Population of the country | Int |
| Unemployment rate | Unemployment rate of a country | float |

**FINER Questions:**

1. Which channel has the highest annual earnings, and what factors contribute to its success, considering the country of origin, the category of the channel, and the viewership metrics?

Reasoning:

Examining the channel that generates the most revenue each year might offer important information about the elements that make it successful. We can spot patterns and trends that might be crucial to the financial success of YouTube channels by taking into account factors like the nation of origin, the type of material, and audience numbers.

1. Which specific channel type for e.g., Music, Entertainment, Education, Sports, or People & Blogs, has the highest number of subscribers globally, and what factors might contribute to the popularity and success of this particular channel type?

Reasoning:

Gaining knowledge about the most popular channel type worldwide and investigating the variables that lead to its success might provide insights into the preferences of the target audience. The success of a specific channel type can be influenced by variables including upload frequency, country of origin, and content category.

1. What insights can be acquired on possible demographic and economic aspects influencing the comparative risk profiles of various places in the context of online content creation, and how can countries help to our understanding of occurrences inside YouTube categories?

Reasoning: This modification highlights the investigation of events inside YouTube categories and makes it clear that the context of online content creation is the search for insights. It keeps the emphasis on the possible impact of economic and demographic variables on the relative risk profiles of various geographical areas.

# What more or different data is needed to address the research questions?

I added two new columns to support my research.

1. years\_since\_creation(till 2023)
2. categoryNum- gave numbers to category column to make it numeric.

# Arranging the Dataset:

The data is well-organized in a tabular format with clear column headers, including details such as subscribers, uploads, rank, video views, country, and more. Every row represents a distinct YouTube channel, providing a standardised framework. The dataset presents information on channels from various nations and content categories. It includes both numerical (such as numbers, profits) and categorical (such as country, category) data. When dealing with big numbers, it is noteworthy to utilise scientific notation. It's crucial to remember that missing value presence needs to be verified. All things considered, the dataset offers a thorough overview for examining trends related to YouTube channels, such as earnings,

# Dropped Columns

Following columns were dropped as they were not needed it seemed to be extra and cant be used for analysis.

1. Title
2. channel\_type
3. video\_views\_rank
4. country\_rank
5. channel\_type\_rank
6. video\_views\_for\_the\_last\_30\_days
7. subscribers\_for\_last\_30\_days
8. created\_date
9. created\_month
10. Gross tertiary education enrollment (%)
11. Latitude
12. Longitude

# Coding:

1. we find mean, median, mode, IQR, minimum, maximum, skewness, lower bound, upper bound using MS EXCEL formulas and made visualizing accordingly.

# Univariate Analysis:

**Subscribers Analysis:**

|  |  |
| --- | --- |
| *Statistics for Subscriber* | |
|  |  |
| Mean | 24456529.52 |
| Standard Error | 822210.5211 |
| Median | 18600000 |
| Mode | 15000000 |
| Standard Deviation | 19439672.04 |
| Sample Variance | 3.77901E+14 |
| Kurtosis | 43.50330477 |
| Skewness | 5.379236208 |
| Range | 232700000 |
| Minimum | 12300000 |
| Maximum | 245000000 |
| Sum | 13671200000 |
| Count | 559 |
| Largest(1) | 245000000 |
| Smallest(1) | 12300000 |

|  |  |
| --- | --- |
| 1st Quartile | 14700000 |
| 2nd Quartile | 18600000 |
| 3rd Quartile | 26500000 |
| IQR | 11800000 |

**From the descriptive statistics for the "Subscriber" variable**:

* The average popularity of YouTube channels in the sample is indicated by the mean number of subscribers, which is roughly 24,456,530.
* The 19,439,672.04 standard deviation indicates a significant variation in the number of subscribers across the channels.
* With a rightward skewness and heavy tails, the kurtosis of 43.503 and skewness of 5.379 signify a severe deviation from a normal distribution.
* With 232,700,000 subscribers, the range—the difference between the maximum and minimum values—is substantial.
* 11,800,000 subscribers is the interquartile range (IQR), a statistical dispersion measure that shows the distribution of the middle 50% of the data.
* There is a minimum of 12,300,000 subscribers and a maximum of 245,000,000.
* Quartile values—the 25th, 50th, and 75th percentiles—offer information about how subscriber counts are distributed.

These numbers indicate that there is a large rightward skew in the subscriber count distribution, indicating that a small number of channels have substantially greater subscriber counts than the rest. The diversity in channel popularity is further supported by the large range and high standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. When assuming a normal distribution, care should be exercised because of the skewed distribution

**video view Analysis:**

|  |  |
| --- | --- |
| *Statistics for Video Views* | |
|  |  |
| Mean | 13003889968 |
| Standard Error | 729494330.9 |
| Median | 8882319696 |
| Mode | #N/A |
| Standard Deviation | 17247566384 |
| Sample Variance | 2.97479E+20 |
| Kurtosis | 61.83775487 |
| Skewness | 6.490467385 |
| Range | 2.28E+11 |
| Minimum | 2634 |
| Maximum | 2.28E+11 |
| Sum | 7.26917E+12 |
| Count | 559 |
| Largest(1) | 2.28E+11 |
| Smallest(1) | 2634 |
| Confidence Level(95.0%) | 1432890598 |

|  |  |
| --- | --- |
| 1st Quartile | 4.94E+09 |
| 2nd Quartile | 8.88E+09 |
| 3rd Quartile | 1.53E+10 |
| IQR | 1.03E+10 |

**From the descriptive statistics for the Video View variable**:

* The average popularity of videos across the YouTube channels in the sample is indicated by the mean number of views, which is roughly 13,003,889,968.
* The 17,247,566,384 standard deviation points to a significant variation in the quantity of video views among channels.
* With a rightward skewness and heavy tails, the kurtosis of 61.838 and skewness of 6.490 signify a severe deviation from a normal distribution.
* With 228,000,000,000 video views, the range—which shows the disparity between the maximum and minimum values—is significant.
* 10,342,875,448 video views is the interquartile range (IQR), a statistical dispersion measure that shows the distribution of the middle 50% of the data.
* There is a minimum of 2,634 views on the video and a maximum of 228,000,000,000 views.
* The 25th, 50th, and 75th percentiles of a quartile offer information about the distribution of video views.

These numbers imply that there is a considerable rightward skew in the distribution of video views, meaning that a small number of videos have a lot more views than the rest. The variety in video popularity is further supported by the large range and significant standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. When assuming a normal distribution, care should be exercised because of the skewed distribution**.**

**Upload variable analysis:**

|  |  |
| --- | --- |
| *Statistics for Uploads* | |
|  |  |
| Mean | 14638.44086 |
| Standard Error | 1867.120989 |
| Median | 1278 |
| Mode | 1 |
| Standard Deviation | 44105.17611 |
| Sample Variance | 1945266560 |
| Kurtosis | 19.44073392 |
| Skewness | 4.25983095 |
| Range | 301307 |
| Minimum | 1 |
| Maximum | 301308 |
| Sum | 8168250 |
| Count | 558 |
| Largest(1) | 301308 |
| Smallest(1) | 1 |
| Confidence Level(95.0%) | 3667.459003 |

|  |  |
| --- | --- |
| 1st Quartile | 435.75 |
| 2nd Quartile | 1296.5 |
| 3rd Quartile | 4187.5 |
| IQR | 3751.75 |

**From the descriptive statistics for the Upload variable**:

* The average number of videos uploaded across the dataset's YouTube channels is 14,638.44, as indicated by the mean number of uploads.
* The 44,105.18 standard deviation indicates that there is significant variation in the quantity of uploads between channels.
* With a rightward skewness and heavy tails, the kurtosis of 19.44 and skewness of 4.26 signify a severe deviation from a normal distribution.
* With 301,307 uploads, the range—the difference between the maximum and smallest values—is sizable.
* 3,751.75 uploads is the interquartile range (IQR), a statistical dispersion measure that shows the distribution of the middle 50% of the data.
* Uploads can have as few as 1 or as many as 301,308.
* The 25th, 50th, and 75th percentiles of a quartile offer information on the upload distribution of upload count.

The results indicate that a small number of channels have substantially more uploads than the bulk, with the upload count distribution being strongly skewed to the right. The variety in upload activity is further supported by the large range and high standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. When assuming a normal distribution, care should be exercised because of the skewed distribution**.**

**Lowest Yearly Earning analysis:**

|  |  |
| --- | --- |
| *Statistics for Lowest Yearly Earnings* | |
|  |  |
| Mean | 625251.9118 |
| Standard Error | 38480.36318 |
| Median | 337900 |
| Mode | 0 |
| Standard Deviation | 909798.1852 |
| Sample Variance | 8.27733E+11 |
| Kurtosis | 15.27747537 |
| Skewness | 3.485908518 |
| Range | 6900000 |
| Minimum | 0 |
| Maximum | 6900000 |
| Sum | 349515818.7 |
| Count | 559 |
| Largest(1) | 6900000 |
| Smallest(1) | 0 |
| Confidence Level(95.0%) | 75584.07005 |

|  |  |
| --- | --- |
| 1st Quartile | 137100 |
| 2nd Quartile | 337900 |
| 3rd Quartile | 702700 |
| IQR | 565600 |

**From the descriptive statistics for the Lowest Yearly Earning variable:**

* The average minimum annual profits for all YouTube channels in the sample are roughly $625,251.91, which is the mean lowest annual earnings.
* The $909,798.19 standard deviation indicates that there is significant variation in the lowest yearly earnings among channels.
* With a rightward skewness and heavy tails, the kurtosis of 15.28 and skewness of 3.49 signify a severe deviation from a normal distribution.
* The lowest yearly earnings in the range, which is the difference between the maximum and minimum figures, are $6,900,000.
* The statistical dispersion measured by the interquartile range (IQR) is $565,600, which represents the distribution of the middle 50% of the data.
* The annual salary ranges from $0 at the lowest to $6,900,000 at the highest.
* The 25th, 50th, and 75th percentiles give information about the distribution of the lowest annual.

The results indicate that a small number of channels have far higher minimum earnings than the others, with the distribution of lowest yearly earnings being strongly biased to the right. The heterogeneity in minimum earnings is further supported by the large range and significant standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. When assuming a normal distribution, care should be exercised because of the skewed distribution.

**Highest Yearly Earning analysis:**

|  |  |
| --- | --- |
| *Statistics for Highest Yearly Earnings* | |
|  |  |
| Mean | 10012263 |
| Standard Error | 616603.1 |
| Median | 5400000 |
| Mode | 2100000 |
| Standard Deviation | 14578458 |
| Sample Variance | 2.13E+14 |
| Kurtosis | 15.1939 |
| Skewness | 3.480117 |
| Range | 1.11E+08 |
| Minimum | 0 |
| Maximum | 1.11E+08 |
| Sum | 5.6E+09 |
| Count | 559 |
| Largest(1) | 1.11E+08 |
| Smallest(1) | 0 |
| Confidence Level(95.0%) | 1211147 |

|  |  |
| --- | --- |
| 1st Quartile | 2200000 |
| 2nd Quartile | 5400000 |
| 3rd Quartile | 11200000 |
| IQR | 9000000 |

**From the descriptive statistics for the Highest Yearly Earning variable:**

* The average maximum annual profits across all YouTube channels in the sample are roughly $10,012,262.97, which is the mean highest yearly earnings.
* The maximum yearly revenues across channels appear to vary widely, as seen by the standard deviation of $14,578,457.81.
* With a rightward skewness and heavy tails, the kurtosis of 15.19 and skewness of 3.48 signify a substantial deviation from a normal distribution.
* The range, which shows the difference between the greatest and lowest

The results indicate that a small number of channels have far greater maximum profits than the remainder, with the distribution of top yearly earnings being strongly biased to the right. The diversity in maximum profits is further supported by the large range and significant standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. When assuming a normal distribution, care should be exercised because of the skewed distribution.

**Population analysis:**

|  |  |
| --- | --- |
| *Statistics for Population* | |
|  |  |
| Mean | 5E+08 |
| Standard Error | 21982340 |
| Median | 3.28E+08 |
| Mode | 3.28E+08 |
| Standard Deviation | 5.2E+08 |
| Sample Variance | 2.7E+17 |
| Kurtosis | -0.83986 |
| Skewness | 0.974138 |
| Range | 1.4E+09 |
| Minimum | 202506 |
| Maximum | 1.4E+09 |
| Sum | 2.8E+11 |
| Count | 559 |
| Largest(1) | 1.4E+09 |
| Smallest(1) | 202506 |
| Confidence Level(95.0%) | 43178249 |

|  |  |
| --- | --- |
| 1st Quartile | 1.08E+08 |
| 2nd Quartile | 3.28E+08 |
| 3rd Quartile | 1.37E+09 |
| IQR | 1.26E+09 |

**From the descriptive statistics for the Population variable:**

* The average population of the nations connected to the YouTube channels in the dataset is roughly 500,229,108, as indicated by the mean population.
* The 519,732,433.2 standard deviation points to a significant range in population sizes between the nations.
* With a minor rightward skewness, the distribution appears to be reasonably symmetric, as indicated by the skewness of 0.97 and kurtosis of -0.84.
* At 1,397,512,494, the range—the difference between the maximum and least values—is quite large.
* 1,397,715,000 is the largest population, and 202,506 is the minimum.
* The 25th, 50th, and 75th percentiles are known as quantile values, and they shed light on population distribution.

The results indicate there is a little rightward skew in the population distribution, with a few nations having substantially larger populations than the others. The variety in population sizes is further supported by the large range and high standard deviation. The middle region of the data's spread can be determined with the use of the quartile values. The minor skewness to the right indicates that more common countries are ones with higher populations.

**Country and Unemployment analysis:**

|  |  |
| --- | --- |
| *Statistics for Unemployment Rate* | |
|  |  |
| Mean | 6.37 |
| Standard Error | 0.60170834 |
| Median | 5.27 |
| Mode | #N/A |
| Standard Deviation | 3.852813258 |
| Sample Variance | 14.84417 |
| Kurtosis | -0.4224573 |
| Skewness | 0.735986505 |
| Range | 13.97 |
| Minimum | 0.75 |
| Maximum | 14.72 |
| Sum | 261.17 |
| Count | 41 |
| Largest(1) | 14.72 |
| Smallest(1) | 0.75 |
| Confidence Level(95.0%) | 1.216097919 |

**From the descriptive statistics for the Country and Unemployment categorized variable:**

* The average unemployment rate for all the nations linked to YouTube channels in the sample is roughly 6.37%, as indicated by the mean.
* The 3.85 standard deviation indicates a moderate degree of variation in the unemployment rates among the nations.
* With a minor rightward skewness, the distribution appears to be reasonably symmetric, as indicated by the skewness of 0.74 and kurtosis of -0.42.
* At 13.97 percentage points, the range—the difference between the maximum and least values—is quite large.
* The unemployment rate ranges from 0.75% at the minimum to 14.72% at the maximum.
* The 25th, 50th, and 75th percentiles of a distribution offer information about how unemployment rates are distributed.

**Detailed graph is visible in excel with all country on X axis.**

In addition, if we look at individual nations, we can see that Jordan has the highest unemployment rate (14.72%) out of all the countries in the dataset. However, with a 0.75% unemployment rate, Thailand has one of the lowest rates in the world. The significant variance in economic conditions across different regions is shown by the disparity in unemployment rates, which highlights the significance of taking into account country-specific elements while comprehending the employment landscape.

Hypothesis Testing:

To answer one of the finer Question, we made the following hypothesis and thus performed testes according to that:

Finer Question: What insights can be acquired on possible demographic and economic aspects influencing the comparative risk profiles of various places in the context of online content creation, and how can countries help to our understanding of occurrences inside YouTube categories?

**Relative Risk**

**Relative Risk Analysis: "Comedy" and "Music" Preferences in India and the United States**

|  |  |  |  |
| --- | --- | --- | --- |
| **Count of category** | **Column Labels** |  |  |
| **Row Labels** | **Comedy** | **Music** | **Grand Total** |
| India | 15 | 26 | 41 |
| United States | 12 | 40 | 52 |
| **Grand Total** | **27** | **66** | **93** |

|  |  |
| --- | --- |
| Country | Relative Risk |
| India | 0.576923077 |
| United States | 0.3 |

After applying the relative risk hypothesis, it can be seen that India, with a computed relative risk of 0.5769, is more likely than the assumed comparison group—presumably the United States—to experience roughly 57.69% of the given occurrence. This suggests that there is a lower likelihood of the event in question happening in India than there is in the US. On the other hand, the US shows an even lower probability, with a relative risk of 0.3, meaning that the event in question is only 30% as likely to occur in the US as it is in the assumed comparison group, which is most likely India.

After taking these relative risk values into account, a thorough conclusion is reached. India and the US both show a lower probability of the mentioned incident in relation to one another. Interestingly, the US shows an even smaller probability than India. These results point to clear disparities in the event's probability between the two nations, emphasising that the occurrence is less likely to happen in the US and India than it is in either country. This emphasises how crucial it is to take into account regional differences when determining the probability of the intended event.

**T-test**

For assessing variations in averages between two independent groups, the t-tests for equal variances and unequal variances are essential instruments. When variances are uneven, the t-test is robust and yields reliable results. On the other hand, the pooled t-test may be less reliable when variances differ since it simplifies calculations by assuming equal variances. The qualities of the data determine which of these tests to use. When comparing nations or YouTube channel categories, in particular, these tests are essential for establishing whether observed differences are statistically significant and supporting well-informed decision-making. Because both tests consider cases with potentially equal and unequal variances, they guarantee a thorough study.

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Unequal Variances | | |
|  |
|  |  |  |  |
|  | *Subscribers* | *Uploads* |  |
| Mean | 24456529.5 | 14648.17889 |  |
| Variance | 3.779E+14 | 1941833428 |  |
| Observations | 559 | 559 |  |
| Hypothesized Mean Difference | 0 |  |  |
| df | 558 |  |  |
| t Stat | 29.7269591 |  |  |
| P(T<=t) one-tail | 2.083E-117 |  |  |
| t Critical one-tail | 1.64758896 |  |  |
| P(T<=t) two-tail | 4.166E-117 |  |  |
| t Critical two-tail | 1.96422445 |  |  |

|  |  |  |
| --- | --- | --- |
| t-Test: Two-Sample Assuming Equal Variances | | |
|  |
|  | *Subscribers* | *Uploads* |  |
| Mean | 24456529.5 | 14648.1789 |  |
| Variance | 3.779E+14 | 1941833428 |  |
| Observations | 559 | 559 |  |
| Pooled Variance | 1.8895E+14 |  |  |
| Hypothesized Mean Difference | 0 |  |  |
| df | 1116 |  |  |
| t Stat | 29.7269591 |  |  |
| P(T<=t) one-tail | 8.185E-144 |  |  |
| t Critical one-tail | 1.64622015 |  |  |
| P(T<=t) two-tail | 1.637E-143 |  |  |
| t Critical two-tail | 1.96209194 |  |  |

The alternative hypothesis (H1) proposes a substantial difference between the mean values of uploads and subscribers, contrary to the null hypothesis (H0), which states that there is no significant difference. The calculated p-value, via a t-test analysis, is surprisingly modest, roughly 1.637 \* 10^-143, well below the traditional significance threshold of 0.05. The null hypothesis is strongly refuted by this incredibly low p-value, which shows that there is, in fact, a significant and statistically significant difference between the mean values of uploads and subscribers. The results confirm that there is a considerable difference between these two variables and provide strong support for the alternative hypothesis. This statistical evidence adds significant insights to our comprehension of the dataset by highlighting the significance of identifying and examining the apparent discrepancy in mean values between subscribers and uploads.

**Chi-Square**

To determine whether data is as expected, we apply the Chi-square test for hypotheses. Comparing the observed values in data to the expected values we would find if the null hypothesis is true is the fundamental notion behind the test. The chi-square statistic indicates the degree of variation between the counts would anticipate in the absence of any association in the population and the observed counts in each table cell. When the chi square test statistic is very tiny, it indicates that the observed and predicted values have a high correlation.





The alternative hypothesis (H1) suggests that there may be a substantial association between "Country" and "Category," contrary to the null hypothesis (H0), which claims there is no significant association. The selected significance level of 0.05 is not met by the calculated p-value of 0.0074. The null hypothesis is thus disproved. The low p-value suggests that there is a statistically significant correlation between "Country" and "Category," which highlights the significance of examining any correlations between these variables in the dataset.

**Odds Ratio**

The probability of an event occurring in one group as opposed to another is evaluated by the odds ratio hypothesis. This metric is frequently employed in logistic regression analysis to investigate the probability of a certain result when a specific variable is present. To put it simply, the odds ratio tells us how the likelihood of an event varies depending on whether a particular element is present or absent. There is no relationship if the odds ratio is 1. A positive association is suggested by values larger than 1, and a negative association is indicated by values smaller than 1. The odds ratio hypothesis provides a nuanced perspective in statistical analysis and is useful in understanding how variables affect the chance of an outcome.

**Odds Ratio Hypothesis: Analyzing the Relationship Between Country (India and United States) and Entertainment Category on YouTube"**

|  |  |
| --- | --- |
| Entertainment | |
| India | 40 |
| United States | 37 |

|  |  |
| --- | --- |
| Odds Ratio | 1.081081081 |

Interpretation: The calculated odds ratio of 1.081 provides evidence against the null hypothesis. The odds ratio being slightly higher for channels from India implies a modest but statistically significant difference in the likelihood of a channel falling into the "Entertainment" category between the two countries. Therefore, the alternative hypothesis is supported, suggesting that there is a significant difference in the odds of a channel being categorized as "Entertainment" based on the country of origin (India or the United States).

**ANOVA- SINGLE FACTOR**

The statistical method known as Analysis of Variance, Single Factor, or ANOVA, is used to determine if the means of three or more groups differ from one another significantly. It allows for the analysis of differences in a single variable between different levels or categories in the context of the report. To ascertain whether the observed variations in means are statistically significant or just the result of chance, the analysis yields an F-statistic and a corresponding p-value. ANOVA Single Factor can be used to investigate and comprehend possible differences between groups, which can help with decision-making and lead to more research into the causes causing observed variations.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Anova: Single Factor |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| SUMMARY |  |  |  |  |  |  |
| *Groups* | *Count* | *Sum* | *Average* | *Variance* |  |  |
| subscribers | 559 | 13671200000 | 24456529.52 | 3.779E+14 |  |  |
| uploads | 559 | 8188332 | 14648.17889 | 1.942E+09 |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
| *Source of Variation* | *SS* | *df* | *MS* | *F* | *P-value* | *F crit* |
| Between Groups | 1.67E+17 | 1 | 1.66975E+17 | 883.6921 | 1.64E-143 | 3.8498048 |
| Within Groups | 2.109E+17 | 1116 | 1.88951E+14 |  |  |  |
|  |  |  |  |  |  |  |
| Total | 3.778E+17 | 1117 |  |  |  |  |

The null hypothesis (H₀) posits no significant difference in the means between the "Subscribers" and "Uploads" groups, while the alternative hypothesis (H₁) proposes a significant difference. The statistical analysis, yielding a remarkably low p-value of 1.637E-143, falls well below the standard significance level of 0.05. Consequently, the null hypothesis is rejected, indicating robust evidence supporting a significant difference in means between "Subscribers" and "Uploads" based on the observed data. In essence, the statistical findings strongly suggest that there is a substantial and noteworthy distinction in the means of these two variables. This insight contributes valuable information to our understanding of the dataset and emphasizes the importance of exploring and acknowledging the observed differences in "Subscribers" and "Uploads."

**Regression 1:**

Regression is a statistical technique that is utilized to simulate the relationship that exists between one or more independent factors and a dependent variable. It is frequently used to forecast the dependent variable's value based on the independent variables' values. Understanding the underlying patterns, quantifying the correlations between variables, and formulating predictions or inferential claims are the objectives.

**Finer Question:** Which channel has the highest annual earnings, and what factors contribute to its success, considering the country of origin, the category of the channel, and the viewership metrics?

Null hypothesis: there is no connection between highest annual earnings and subscribers.

Alternative hypothesis: relationship between highest annual earnings and subscribers



According to the regression study, there is a significant positive correlation (Multiple R = 0.71) between the annual earnings and the number of subscribers. The subscribers account for around half of the earnings fluctuation (R Square = 0.50). The low standard error (9780003.36) indicates that there is not much variation between earnings as expected and actual. The model has 995 observations, which is a strong support. The model's statistical significance is highlighted by a strong F-statistic (985.24) and an incredibly low p-value (8.64E-151).

When looking at coefficients, the intercept (2809445.12) is essential to the model but isn't realistically significant at zero subscribers. According to the subscribers' coefficient (18.51), a large rise in annual revenue is often associated with each new subscriber. With a p-value of 8.64E-151, this association is extremely statistically significant.

Finally, the model concludes that half of the earnings fluctuation may be explained by the robust prediction of annual earnings based on the number of subscribers. This provides important new information about the dynamics of success for content creators on the network.

**Regression 2:**

**Finer Question**: Which specific channel type for e.g., Music, Entertainment, Education, Sports, or People & Blogs, has the highest number of subscribers globally, and what factors might contribute to the popularity and success of this particular channel type?

**Channel Type with the Highest Subscribers:** With the greatest number of subscribers (64441018), channel type 5 (Entertainment) appears to be the most well-liked worldwide.

**Factors Contributing to Popularity and Success:** The positive coefficients show that there is a positive association between subscriptions and population. Subscriptions rise in tandem with population growth.

The p-value linked to the population coefficient indicates the importance of the association. A p-value of less than a predetermined cutoff (usually 0.05) indicates that the link may be statistically significant.

It's crucial to understand that even in cases when a statistical relationship exists, causality is not implied by correlation. This model does not account for other variables that can affect the number of subscribers.

**In conclusion:** The model indicates that population is a statistically significant factor determining the number of subscribers, even if it only accounts for a small portion of the variability. The category with the most subscribers is clearly channel type 5 (Entertainment), although further research is required to fully comprehend the complex elements influencing each channel type's performance.

|  |  |  |
| --- | --- | --- |
| **CategoryNum** | **Sum of subscribers** | **Sum of Population** |
| 1 | 20301808 | 15045573037 |
| 2 | 800000 | 222329946 |
| 3 | 19800256 | 32253105665 |
| 4 | 10500000 | 27376884878 |
| 5 | 64441018 | 95721572163 |
| 6 | 18202965 | 20938827955 |
| 7 | 12900039 | 15513569223 |
| 8 | 3400003 | 14802106657 |
| 9 | 800000 | 2732835508 |
| 10 | 24311224 | 70674991593 |
| 11 | 5630000 | 17957081993 |
| 12 | 500000 | 540798940 |
| 13 | 34006707 | 39720610043 |
| 14 | 1300005 | 723313451 |
| 15 | 3100042 | 7875624638 |
| 16 | 6500002 | 8290534950 |
| 17 | 2600000 | 2049086436 |
| 18 | 600000 | 2732835508 |
| 19 | 0 | 126014024 |

The percentage of the dependent variable's variation (the total number of subscribers) that can be predicted from the independent variable (CategoryNum) is denoted by R². With an R2 of 0.1216, the variation in CategoryNum may account for around 12.16% of the variability in the sum of subscribers.

**Discussing overall study results:**

1. What do the data and analyses tell you and your audience, in relationship to the problems or opportunity or question you are trying to address?

Ans: The analyses offer insightful information about the variables impacting YouTube channel success, with a special emphasis on category preferences, economic issues, and demographics. The examination of relative risk revealed differences in the likelihood of events between the United States and India. Odds ratio analyses, chi-square tests, and t-tests helped to clarify the statistical significance of different associations. The results of the ANOVA showed that the means of the various groups differed significantly.

1. Did your analyses answer your questions? If so, how? If not, why not?

Ans: My research's issues have been addressed by the analyses, which also offered statistical support and new perspectives on the connections between various factors.

Every statistical test and regression analysis added to our understanding of the variables affecting the production of online content.

1. Did you run into issues with your data? Were there missing elements, or problems with data quality (with reference to cleaning discussed above, or discovered later)?

Ans: Although data cleansing was not specifically addressed, it is crucial to preserve data quality. Resolving any problems, such as incomplete data or quality issues, guarantees the validity of the findings. To avoid data loss and changing the meaning only required data was cleaned.

1. Did you discover new questions that could or should be investigated, with this or other data?

Ans: The in-depth investigations might have raised more queries, especially in consideration of the complex variables affecting YouTube channel success. Subsequent investigation can focus on the unique characteristics of successful channels in every category.

**Conclusion**:

My investigations produced convincing responses to my main study questions by revealing the complexities of online content generation. The relative risk study highlighted the need to take regional variations into consideration by illuminating subtle variations in the probabilities of content-related incidents between the United States and India. Subscription counts, uploads, and preferred categories varied significantly, as demonstrated by t-tests and chi-square studies.

Regression analysis revealed important insights. The significant positive relationship between yearly earnings and subscribers highlighted how important a subscriber base is to the success of a channel. Moreover, an investigation on the impact of channel classifications on the quantity of subscribers highlighted the importance of population as a contributing factor.

There are serious implications. It becomes critical for content providers to understand regional differences and the significant influence that subscriber numbers has on income. The process of strategically designing content begins with the identification of influential characteristics within channel categories. These findings provide us with practical insights to help us navigate through online content, making educated decisions and optimizing our chances of success.

1. Which channel has the highest annual earnings, and what factors contribute to its success, considering the country of origin, the category of the channel, and the viewership metrics?

Answer: Relationship between highest annual earnings and subscribers exists

Music channel has highest number of yearly earnings

1. Which specific channel type for e.g., Music, Entertainment, Education, Sports, or People & Blogs, has the highest number of subscribers globally, and what factors might contribute to the popularity and success of this particular channel type?

Answer: With the greatest number of subscribers (64441018), channel type 5 (Entertainment) appears to be the most well-liked worldwide.

1. What insights can be acquired on possible demographic and economic aspects influencing the comparative risk profiles of various places in the context of online content creation, and how can countries help to our understanding of occurrences inside YouTube categories?

Answer: India and USA show low probability in terms on Music and comedy preferences

**Tracking Demonstration:**

1. Data Collection:

Sources: <https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data>

provided the dataset used in this analysis.

1. Descriptive Statistics:

The variables chosen are: Important factors were Country vs Unemployment rate, subscribers, video views, uploads, lowest year earning, highest year earning, population which were essential to understanding the dynamics of the content.

Characteristic Measures: To identify key trends and subtleties in probability distributions, the measures of mean, variance, and relative risk were calculated.

1. Data Cleaning and Categorization:

Data Handled: Null values were handled and columns which were not required were removed.

Classification: categories country and employment were categorized

1. Hypothesis Formulation and Testing:

Formulated hypotheses and chose techniques like relative risk, t-test, anova, odds ratio and chi-square

1. Regression Analyses:

Two regression analysis were performed based on finer questions

1. Visualization was made for every result including univariate, hypothesis testing and inferential techniques.