Final Project

Daxeshkumar Patel

MGT 2148 – Business Analytics and Decision Making

Mohammad Raahemi

11th August 2023

|  |
| --- |
| DESCRIPTION PAGE NO. |
| INTRODUCTION……………………………………………………………………..3  Scope of Analysis………………………………………………………... 3  Visual Insight………………………………………………………………4 |
| DATA GATHERING AND CLEANING FOR ANALYSIS………………………...5 |
| MEAN, MEDIAN AND MODE OF DATA………………………………………….6   1. Mean……………………………………………………………………...6 2. Median……………………………………………………………...……6 3. Mode……………………………………………………………………...7 4. Analytical Conclusions of Mean, Median And Mode………………...7 |
| VISUAL ANALYSIS BY CHART AND DIAGRAM………………………………9   1. Pie Chart Analysis (Subscribers by Category)…………………….....9 2. Bar Chart Analysis (Yearly earning of top 10 YouTube)…………….11 3. Line Chart Analysis (Views of Different Years)………………………12 4. Tree-map Analysis (Video Uploaded by channels)………………….13 |
| HYPOTHESIS TESTING…………………………………………………………...15   1. Hypothesis - Two Tail……………………………………………………15 2. Hypothesis - One Tail right……………………………………………..17 3. Hypothesis - One Tail left………………………………………………19 |
| REGRESSION………………………………………………………………………..22   1. Simple Linear Regression……………………………………………….22 2. Multiple Linear Regression………………………………………………25   REFERENCE………………………………………………………………………….28 |
|  |

**INTRODUCTION**

Data-driven decision-making is becoming a crucial strategy for businesses to achieve a competitive edge and track the complexity of the market in today’s competitive world. Businesses can understand their consumers in better way and improves their operations by being able to get useful insights from data. **This research explores the world of business analysis and market judgement, concentrating on the examination of a sizable YouTube dataset of year 2022 (uploaded in September 2022).**

This report focused on the application of Business Analysis principles through a meticulous examination of a rich **YouTube dataset, encompassing** essential information about YouTube channels, their **subscribers, viewership, uploads, geographical origins, and the years of their creation**. It is also a valuable resource for **understanding the dynamics of content creation, audience engagement, and the potential impact on market decisions.** In this report we have included study of statistical measures like **mean, median, and mode, graphical analysis of data** as well as **advanced analysis techniques** like **hypothesis testing and regression analysis**.

**SCOPE OF ANALYSIS**

Our investigation spans a broad spectrum of analytical techniques, each contributing to a comprehensive understanding of the dataset's implications. **The statistical measures of mean, median, and mode shed light on central tendencies within the data, providing a foundational understanding of its distribution. Furthermore, hypothesis testing facilitates the validation or rejection of claims about the dataset, while regression analysis enables us to uncover relationships and predictive insights.**

**VISUAL INSIGHTS**

To enhance our understanding and present our findings effectively, we use the power of data visualization. Through the creation of various charts and graphs, we aim to visualize trends, correlations, and patterns within the dataset. These visual representations serve as **powerful tools for conveying complex information** in an accessible manner.

**DATA GATHERING AND CLEANING FOR ANALYSIS: -**

We have started with Global YouTube statistics dataset of year 2022 from [**https://www.kaggle.com**](https://www.kaggle.com)which is reliable source of data. The dataset encompassed insights from the top **995 YouTube channels worldwide**.

During the initial stages of **data processing, certain inconsistencies** came to notice. Specifically, we identified instances where **some data cells were lacking information, others contained unrelated data, and a few even presented negative counts** that are not possible in practical terms. In order to resolve these inconsistencies, we have used Microsoft Excel as a tool to **accurately cleanse and refine the original dataset.**

Through a thorough process of error identification and removal, we successfully **transformed the original dataset into a refined dataset**. As a result, we were able to curate a dataset consisting of data from **785 YouTube channels** which became popular for us and could lead us to precise analysis results.

**MEAN, MEDIAN AND MODE OF DATA**

1. **Mean: -**

The "Mean" of data is referred to as the "average," is a measure of central trend that represents the average of a set of values. It's calculated by adding up all the values in a dataset and then dividing that sum by the total number of values of the dataset.

Mathematically, the mean (μ) of a dataset with n values (x1, x2, x3,...,xn) is calculated as:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CREATER\_YEAR** | **SUBSCRIBERS** | **VIDEO VIEWS** | **UPLOADS** | **POPULATION** | **YEARLY EARNING (USD)** |
| **MEAN** | 2013 | 22779745 | 11552851491 | 11372 | 443745787 | 522707 |

1. **Median: -**

Median is the **middle value** in a dataset when the values are arranged in ascending or descending order.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CREATER\_YEAR** | **SUBSCRIBERS** | **VIDEO VIEWS** | **UPLOADS** | **POPULATION** | **YEARLY EARNING (USD)** |
| **MEDIAN** | 2013 | 17500000 | 7776706184 | 1021 | 328239523 | 243700 |

1. **Mode: -**

Mode is the **data point** that occurs with the **highest frequency**.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **CREATER\_YEAR** | **SUBSCRIBERS** | **VIDEO VIEWS** | **UPLOADS** | **POPULATION** | **YEARLY EARNING (USD)** |
| **MODE** | 2014 | 12500000 | #N/A | 147 | 328239523 | 1100000 |

1. **ANALYTICAL CONCLUSIONS OF MEAN, MEDIAN AND MODE:**
2. **CREATER\_YEAR:**

* **Mode of 2014** indicates that among the top 785 youtubers **maximum creators** came on platform **during the year 2014 are successful**.

1. **SUBSCRIBERS:**

* The **wide gap between the mean** (22,779,745) **and median** (17,500,000) subscriber counts **indicates** the **presence of creators with exceptionally high subscriber counts, or outliers.**
* **The mode** subscriber count at **12,500,000** could suggest that a from top You Tubers, **majority of creators** falls within a **similar subscriber range.**

1. **VIDEO VIEWS:**

* The lack of a mode value for video views suggests that **no single value occurs most frequently**, indicating a **diverse range of video views** on platform.
* The **considerable difference between the mean and median** suggests that a **few creators with extremely high view counts** are driving up the mean or in other terms **You Tube business**.

1. **UPLOADS:**

* The **mean and median values for uploads** are **considerably higher than the mode**, indicating that a **few creators** **with a high number of uploads** might be **influencing the channel**.
* The **low mode** value suggests that **most creators** in the **dataset have a relatively moderate number of uploads**.

1. **POPULATION:**

* The **identical median and mode population** values indicate that **one population value occurs most frequently,** indicates that the dataset is **focused on creators** from a specific regions.

1. **YEARLY EARNING (USD):**

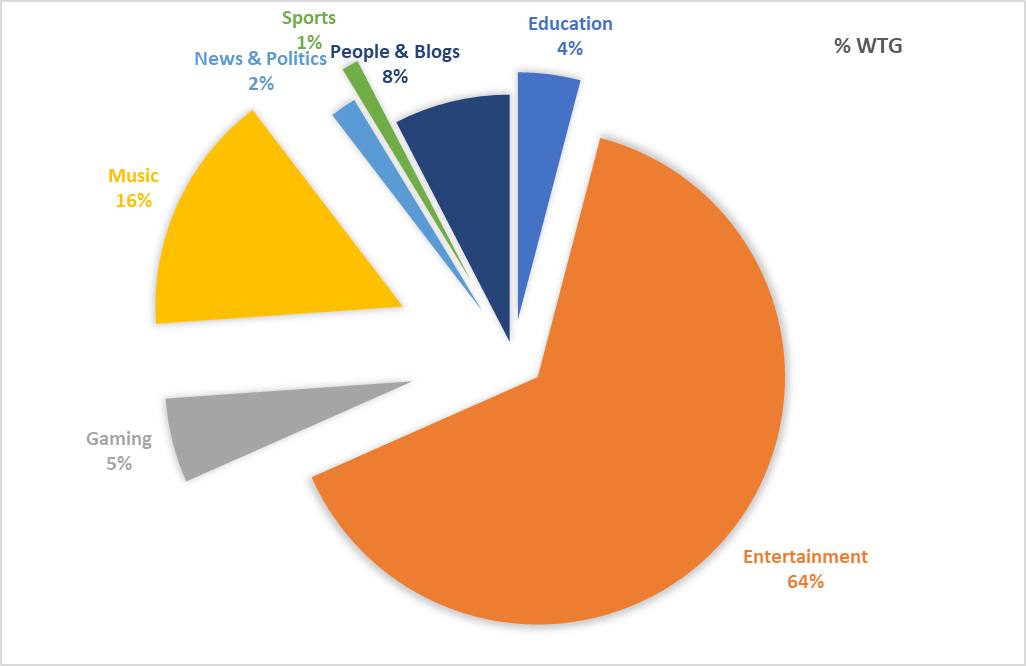
The **considerable difference between** the **mean and median** **earnings** indicates **potential outliers with high earnings that are influencing the mean**.

The **mode value for earnings at $1,100,000** represent **creators with earnings around this value**.

This data analysis suggests that the dataset contains a mix of creators. Further analysis could examine relationships between variables, explain correlations, and identify trends within the top YouTube creator landscape.

**VISAL ANALYSIS BY CHART AND DIAGRAM**

1. **Pie Chart Analysis for Subscribers Distribution by Category of content:**

****

This visualization gives you a quick overview of how subscribers are proportionally distributed among the various content categories on YouTube.

1. **CONTENT DOMINANCE:**

* The **Entertainment category** commands the largest share of subscribers, **accounting for 64.30% of the total**.
* This dominance suggests that **entertainment-related content** attracts the **highest portion.**

1. **Engagement:**

* The **Music category** holds the **second-largest share** of subscribers, constituting **15.72%.**
* This demonstrates a **diverse range of content**.

1. **Education's Impact:**

* The Education category captures a **notable portion with 4.07%** of subscribers.
* This suggests that educational **content is relevant and engaging to a significant portion** for business.

1. **Niche Categories:**

* The **Gaming and People & Blogs** categories contribute same which is approximately **13% of subscribers**.
* These **niche categories indicate** the **presence of dedicated audiences** in specific types of content.

1. **Limited Engagement:**

* The **Sports and News & Politics** categories hold **around 1%**.
* This suggests that **these categories have a relatively limited audience**.

1. **Marketing Insights:**

* YouTube can take benefit of the dominance of the **Entertainment and Music categories** for **targeted marketing campaigns**.
* **Niche categories** can also be valuable for **reaching engaged, specialized audiences.**

1. **Potential Growth Areas:**

* Categories with **smaller subscriber** shares, such as **Sports and News & Politics,** could **represent growth opportunities** for content creators targeting underserved audiences.
* This **analysis of the pie chart** provides insights into **how subscribers are distributed across different YouTube content** categories. These insights can guide businesses in **decision making** about **content creation, marketing strategies, and audience engagement.**

1. **Bar Chart analysis for Yearly earning of top 10 YouTube Channels:**

This visualization gives you a quick overview of top 10 YouTube channels and their earnings earning which allowing to assess the differences in financial performance

1. **Earnings Disparities:**

* The bar chart highlights substantial **disparities in yearly earnings** among the top 10 YouTube channels.
* **T-Series stands** out with the highest earnings of 6,800,000 USD, showcasing **its dominance** within the platform and even being in top 10 list **PewDiePie** is earning 117,600 USD which is **significantly less** than the other channels.

1. **Monetization Strategies:**

* The varying earnings highlight the diverse monetization strategies that channels employ, including ad revenue, partnerships, merchandise sales, and more.

1. **Corporate vs. Independent Earnings:**

* High earnings of **Zee Music Company** and **WWE** suggest the **revenue potential** for established corporate entities on YouTube.
* Independent creators like **MrBeast and PewDiePie**, however, can also **compete significantly in terms of earnings.**

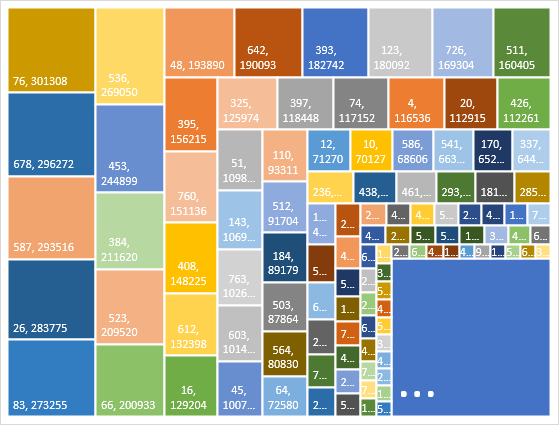
1. **Line Chart Analysis For Views Earned Of YouTube Channels Started In Different Years:**
2. **Growth of Viewers and Audience Reach: -**

* YouTube channels started in the year 2006 and 2007 are getting the maximum views in other words they have best reach to audience.
* YouTube should select those channels to run the advertisement for getting the better viewer reach.

1. **Average Performing Creators: -**

* For the creator who has started in year 2011 and 2014 are getting average 800 billion views which should be second highest after year 2006-07, which makes them ideal for running the advertisement.
* Further YouTube should focus on those channels which started in year 2011 and 2014 since they have a higher possibility of being future leaders in terms of view generation.

1. **Tree-map Analysis for Video Uploaded by channels:**



1. **Ranking vs. Uploads:**

* There is **no correlation between a channel's ranking and the number of videos uploaded.** It is not mandatory to upload the maximum number of videos to get successful on YouTube.

1. **Distribution of Uploads:**

* We can visualize the distribution of video uploads across the channels.
* Majority of channels have a relatively low number of uploads, while a smaller number of channels have a significantly higher number of uploads.

1. **Outliers for Data:**

* Outliers can be easily detected by the Tree-Map Analysis which helps us to do **further analysis based** on finding the **higher and lower values of data.**

**HYPOTHESIS TESTING: -**

1. **HYPOTHESIS - TWO TAIL**

Data for Two Tail Hypothesis Testing: -

We have filtered data for two countries (India and US), taken random sample of 10 YouTube Channel and decided to do hypothesis testing for Yearly Earning

**Data Samples: -**

**For INDIA**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Yearly Earning (USD)** |
| 18 | ChuChu TV Nursery Rhymes & Kids Songs | IN | Education | 1300000 |
| 76 | ABP NEWS | IN | People & Blogs | 801200 |
| 165 | BB Ki Vines | IN | Entertainment | 51800 |
| 300 | Kiddiestv Hindi - Nursery Rhymes & Kids Songs | IN | Education | 340300 |
| 445 | Bollywood Classics | IN | Entertainment | 454400 |
| 503 | Mazhavil Manorama | IN | Entertainment | 1100000 |
| 534 | Sony Music South | IN | Music | 726600 |
| 643 | Kids TV India Hindi Nursery Rhymes | IN | Shows | 236000 |
| 705 | Lokdhun Punjabi | IN | Music | 103300 |
| 785 | Make Joke Of | IN | Comedy | 72000 |

**For US**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Yearly Earning (USD)** |
| 28 | EminemMusic | US | Music | 780600 |
| 78 | Bruno Mars | US | Music | 525700 |
| 121 | Diana and Roma ESP | US | nan | 309000 |
| 163 | Alan Chikin Chow | US | Comedy | 3100000 |
| 201 | DJ Snake | US | Music | 189900 |
| 240 | Sia | US | Music | 285800 |
| 422 | CoryxKenshin | US | Gaming | 315300 |
| 585 | Telemundo | US | Entertainment | 144900 |
| 656 | DALLMYD | US | Sports | 92000 |
| 741 | Future | US | Music | 1400 |

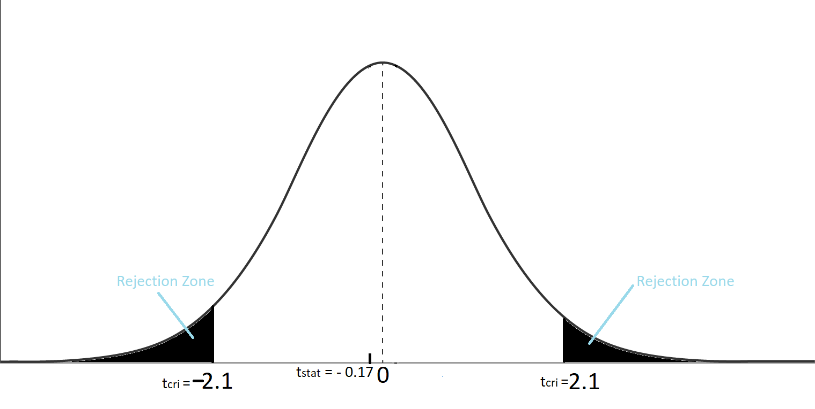
**Assumptions for Two Tail Hypothesis Test: -**

* Standard deviation is not given hence we are using the **T distribution table**
* We are also assuming that the **variances are equal**.
* Assuming the **confidence level as 95%**
* We are taking **random samples of 10 entries** for **India and US**
* Earning of India Youtubers
* Earning of US Youtubers
* For the purposes of the project, we are assuming 2 hypothesis conditions: Null () and Alternate ()
* Null : U1=U2 (Earnings of Indian YouTubers and US Youtubers are equal)
* Alternate : U1≠U2 (Earnings of Indian YouTubers and US Youtubers are not equal)

**Result Output: -**

|  |  |  |
| --- | --- | --- |
|  | *INDIA* | *US* |
| Mean | 518560 | 574460 |
| Variance | 1.97482E+11 | 8.37528E+11 |
| Observations | 10 | 10 |
| Pooled Variance | 5.17505E+11 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 18 |  |
| t Stat | -0.173755829 |  |
| P(T<=t) one-tail | 0.431998262 |  |
| t Critical one-tail | 1.734063607 |  |
| P(T<=t) two-tail | 0.863996524 |  |
| t Critical two-tail | 2.10092204 |  |

**Key Take aways of Result Table: -**

**Mean and Variance:** The mean for India is 518,560 and for the US is 574,460.

**Observations:** Number of Observation for India and US both are 10.

**Degrees of Freedom (df):** 18 (20-2)

tStat**:** The calculated t-statistic is -0.17

tCrit for Two Tail test: 2.100

Value of tStat 0.17) does not fallwithin a rejection zone which indicates that we do not have enough evidence to reject the null hypothesis.

**Therefore, we can say that we do not have enough evidence to conclude that there is a statistically significant difference in the YouTube Earnings between India and the US Youtubers.**

1. **HYPOTHESIS - ONE TAIL RIGHT**

Data for One Tail Right Hypothesis Testing: -

We have filtered data for year 2014 and 2012, taken random sample of 6 YouTube Channel and decided to do hypothesis testing for Yearly Uploads.

**Data Samples: -**

**For Year 2014**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Video Uploads** |
| 219 | Miss Katy | GB | People & Blogs | 1044 |
| 224 | Wave Music Bhojpuri | IN | Music | 22578 |
| 200 | netd mï¿½ï¿ | TR | Music | 23491 |
| 303 | MiawAug | ID | Gaming | 3713 |
| 368 | SonyMusicSouthVEVO | US | Music | 5692 |
| 508 | Sony LIV | IN | Entertainment | 11099 |

**For Year 2012**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Video Uploads** |
| 13 | BANGTANTV | KR | Music | 2281 |
| 113 | Worldwide Records Bhojpuri | IN | Music | 6518 |
| 137 | Enaldinho | BR | Comedy | 3654 |
| 383 | Porta dos Fundos | BR | Entertainment | 2240 |
| 568 | Jesser | US | Gaming | 1244 |
| 621 | AM3NlC | BR | Gaming | 5757 |
| 741 | Future | US | Music | 1400 |

**Assumptions for One Tail Right Hypothesis Test: -**

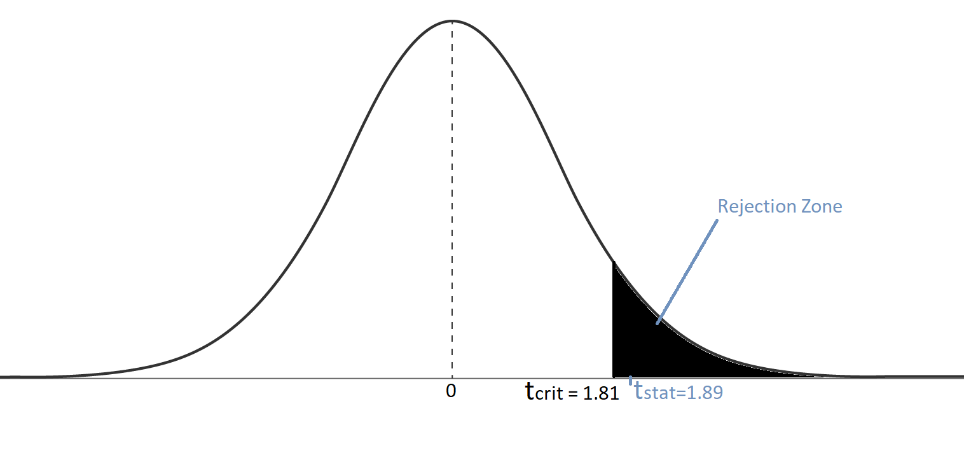
* Standard deviation is not given hence we are using the **T distribution table**
* We are also assuming that the **variances are equal**.
* Assuming the **confidence level as 95%**
* We are taking **random samples of 6 entries** for **2014 and 2012 upload data**
* No. of uploads for year 2014
* No. of uploads for year 2012
* For the purposes of the project, we are assuming 2 hypothesis conditions: Null () and Alternate ()
* Null : U1=U2 (Video Uploads for year 2014 are equal as of 2012)
* Alternate : U1>U2 (Video Uploads for year 2014 are greater than year 2012)

**Result Output: -**

|  |  |  |
| --- | --- | --- |
|  | *Year 2014* | *Year 2012* |
| Mean | 11269.5 | 3615.667 |
| Variance | 94009270.7 | 4461783 |
| Observations | 6 | 6 |
| Pooled Variance | 49235526.68 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 10 |  |
| t Stat | 1.889297432 |  |
| P(T<=t) one-tail | 0.044081972 |  |
| t Critical one-tail | 1.812461123 |  |
| P(T<=t) two-tail | 0.088163943 |  |
| t Critical two-tail | 2.228138852 |  |

**Key Take aways of Result Table: -**

**Mean and Variance:** The mean for video uploaded during the year 2014 is 11,269 and for the year 2012 it is 3,615.

**Observations:** Number of Observation for year 2014 and 2012 are 6.

**Degrees of Freedom (df):** 10 (12-2)

tStat**:** The calculated t-statistic is 1.89

tCrit for One Tail test:1.81

Value of tStat : 1.89 > tCrit :1.81, hence value of tStat fallwithin a rejection zone which indicates that we have enough evidence to reject the null hypothesis.

**Therefore, we can say that based on hypothesis testing, there is a significant difference in Video uploads between the years 2014 and 2012.**

1. **Hypothesis - One Tail Left**

Data for One Tail Left Hypothesis Testing: -

We have filtered data for Entertainment and Educational YouTube channels and taken random sample of 10 YouTube Channel and decided to do hypothesis testing for Channel Wide Views.

**Data Samples: -**

**For Year Entertainment**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Video View** |
| 8 | Vlad and Niki | US | Entertainment | 77180169894 |
| 90 | Jess No Limit | ID | Entertainment | 3963007415 |
| 154 | Dushyant kukreja | IN | Entertainment | 17318452893 |
| 229 | Tsuriki Show | DE | Entertainment | 31494513067 |
| 294 | SlivkiShow | UA | Entertainment | 4707412332 |
| 397 | AMARINTV | TH | Entertainment | 13043561912 |
| 528 | Little Angel | BR | Entertainment | 8897705695 |
| 601 | RS 1313 SHORTS | IN | Entertainment | 8011977288 |
| 684 | Vogue | US | Entertainment | 4177184071 |
| 764 | Flowers Comedy | IN | Entertainment | 11691081301 |
| 8 | Vlad and Niki | US | Entertainment | 77180169894 |

**For Year 2012**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Video View** |
| 3 | Cocomelon - Nursery Rhymes | US | Education | 164000000000 |
| 32 | Infobells - Hindi | IN | Education | 30516172739 |
| 62 | Super Simple Songs - Kids Songs | CA | Education | 44900897958 |
| 101 | BabyBus - Kids Songs and Cartoons | US | Education | 26355088167 |
| 234 | BabyBus - Canciones Infantiles & Videos | MX | Education | 13206471140 |
| 339 | Little Angel | MX | Education | 9813245108 |
| 461 | wifistudy | IN | Education | 3018032423 |
| 514 | Daftar Populer | ID | Education | 2143140898 |
| 637 | Veritasium | US | Education | 2165885634 |
| 735 | Right to Shiksha | IN | Education | 7876740921 |

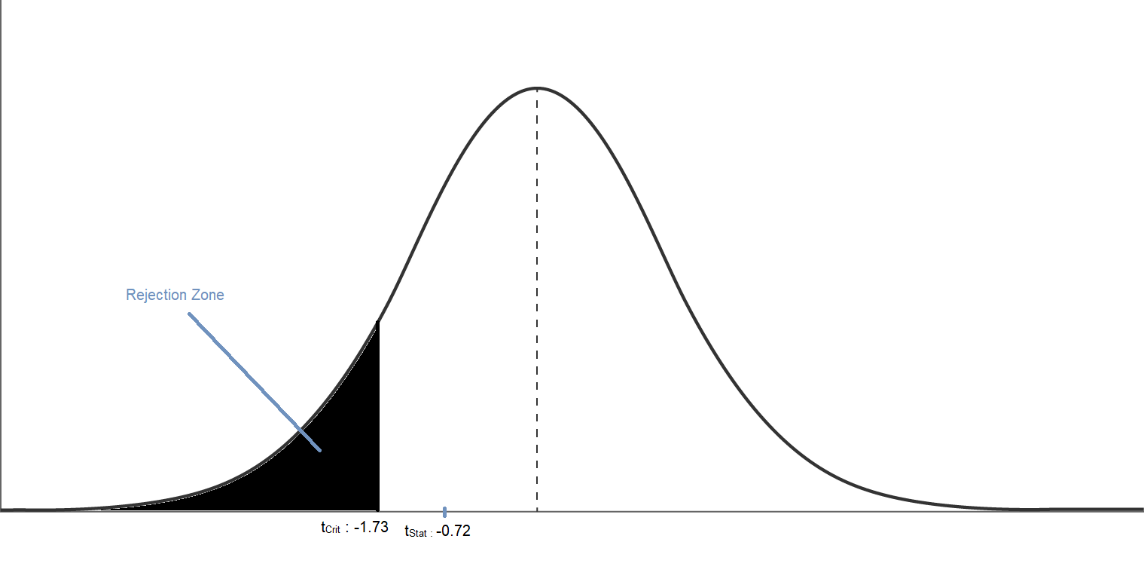
**Assumptions for One Tail Left Hypothesis Test: -**

* Standard deviation is not given hence we are using the **T distribution table**
* We are also assuming that the **variances are equal**.
* Assuming the **confidence level as 95%**
* We are taking **random samples of 10 entries** for **Entertainment** **and Educational Video View data**
* No. of Video Views for Entertainment Category
* No. of Video Views for Education Category
* For the purposes of the project, we are assuming 2 hypothesis conditions: Null () and Alternate ()
* Null : U1=U2 (Video Views for year Entertainment are equal as of Education)
* Alternate : U1<U2 (Video Views for year Entertainment are less than Education)

**Result Output: -**

|  |  |  |
| --- | --- | --- |
|  | *Entertainment* | *Education* |
| Mean | 18048506587 | 30399567499 |
| Variance | 4.99624E+20 | 2.4053E+21 |
| Observations | 10 | 10 |
| Pooled Variance | 1.45246E+21 |  |
| Hypothesized Mean Difference | 0 |  |
| df | 18 |  |
| t Stat | -0.72466393 |  |
| P(T<=t) one-tail | 0.238986559 |  |
| t Critical one-tail | 1.734063607 |  |
| P(T<=t) two-tail | 0.477973118 |  |
| t Critical two-tail | 2.10092204 |  |

**Key Take aways of Result Table: -**

**Mean and Variance:** The mean for video view for the category Education and Entertainment is given

**Observations:** Number of Observation for year Education and Entertainment are10.

**Degrees of Freedom (df):** 18 (20-2)

tStat**:** The calculated t-statistic is -0.72

tCrit for One Tail test: -1.73

Value of tCrit : -1.73 < tStat: -0.72. Hence the value of tStat does not fallwithin a rejection zone which indicates that we don’t have enough evidence to reject the null hypothesis.

**Therefore, we can say that based on the provided data, there is a no significant difference in Video Views between Entertainment and Education.**

**REGRESSION:**

* Regression gives us the relationship between interval variables.
* Values of one variable (Dependent Variable) based on other (Independent Variable) can be predicted by Regression Analysis.

**There are two types of regression:**

1. Simple Linear Regression: One Variable and One Independent Variable
2. Multiple Linear Regression: One Variable and Multiple Independent Variables
3. **Simple Linear Regression:**

Here we are taking the example of Canadian Youtubers considering their yearly earnings are dependent on their Subscriber count.

Dataset sample from Canadian youtuber is taken as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Country** | **Category** | **Subscribers** | **Yearly Earning (USD)** |
| 14 | Justin Bieber | CA | Music | 71600000 | 529000 |
| 62 | Super Simple Songs | CA | Education | 39200000 | 1700000 |
| 105 | The Weeknd | CA | Music | 32600000 | 1100000 |
| 133 | Shawn Mendes | CA | Music | 30000000 | 266200 |
| 171 | VanossGaming | CA | Gaming | 25800000 | 164800 |
| 183 | WatchMojo.com | CA | Music | 25100000 | 147000 |
| 480 | Linus Tech Tips | CA | Science & Technology | 15600000 | 264700 |
| 590 | Lilly Singh | CA | Comedy | 14500000 | 16000 |
| 639 | Hacksmith Industries | CA | Science & Technology | 13800000 | 251800 |
| 662 | Typical Gamer | CA | Gaming | 13500000 | 97700 |
| 691 | AzzyLand | CA | Gaming | 13300000 | 16200 |
| 733 | MrSuicideSheep | CA | Entertainment | 12800000 | 54400 |
| 773 | Heidi and Zidane HZHtube | CA | Entertainment | 12400000 | 28800 |

**Outputs:**

We have performed simple linear regression on above data and got below results:

Equation for Liner regression: y = - 35505x + 379991, R² = 0.5774

**SCATTER DIAGRAM OF REGRESSION MODEL:**



**Key Takeaways of Simple Linear Regression Analysis:**

379991

- 35505 (Slop is negative)

**Standard error**

87863.20

Relatively speaking **our linear regression model** of **YouTube earning as a function of Subscriber does not fit well.**

**Test to determine if there is a linear relationship between the YouTube earning and Subscribe (at 5% significant number):**

**: = 0**

**: 0**

We have done one tail right hypothesis testing for solution.

if the null hypothesis is true, no linear relationship exists.

tStat**:** The calculated t-statistic is 4.70

tCrit: for One Tail right test: 1.73

**There is enough evidence to infer that a linear relationship exists between subscriber count and YouTube earnings.**

**Coefficient of Determinant (R2)**

To measure the strength of the relationship we use coefficient of determinant (R2)

R2 = 0.710680494

**Which means 71.06% variation YouTube earning is expected by the variance in Subscriber count and remaining 28.94 is unexpected, due to error.**

Coefficient of determinant does not have a critical value that enable us to draw conclusion however higher the R2 better the model fits data.

If it R2 is 1 then it would be considered as perfect match and if R2 is 0 then it would be considered as no linear relation.

In our case R2 is near to 1 hence we can say that model is better fit for the data taken.

1. **Multiple Regression: -**

The simple linear regression model is used to determine the relationship. However in Multiple Regression method we use multiple independent variable to determine the single dependent variable.

Here we are taking the sample from dataset of top 20 Youtubers **considering their yearly earnings are dependent** on their Subscriber count, Video Views and number of videos uploaded.

Dataset sample from top 20 Youtuber is taken as below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Rank** | **Youtuber** | **Subscribers** | **Video Views** | **Uploads** | **Yearly Earning (USD)** |
| 1 | T-Series | 245000000 | 228000000000 | 20082 | 6800000 |
| 2 | MrBeast | 166000000 | 28368841870 | 741 | 4000000 |
| 3 | Cocomelon - Nursery Rhymes | 162000000 | 164000000000 | 966 | 5900000 |
| 4 | SET India | 159000000 | 148000000000 | 116536 | 5500000 |
| 5 | ýýý Kids Diana Show | 112000000 | 93247040539 | 1111 | 2200000 |
| 6 | PewDiePie | 111000000 | 29058044447 | 4716 | 117600 |
| 7 | Like Nastya | 106000000 | 90479060027 | 493 | 146800 |
| 8 | Vlad and Niki | 98900000 | 77180169894 | 574 | 1700000 |
| 9 | Zee Music Company | 96700000 | 57856289381 | 8548 | 2400000 |
| 10 | WWE | 96000000 | 77428473662 | 70127 | 2100000 |
| 11 | BLACKPINK | 89800000 | 32144597566 | 543 | 1500000 |
| 12 | Sony SAB | 83000000 | 101000000000 | 71270 | 5000000 |
| 13 | BANGTANTV | 75600000 | 20826993957 | 2281 | 504900 |
| 14 | Justin Bieber | 71600000 | 30608119724 | 249 | 529000 |
| 15 | HYBE LABELS | 71300000 | 28634566938 | 1337 | 1800000 |
| 16 | Zee TV | 70500000 | 73139054467 | 129204 | 5100000 |
| 17 | Pinkfong Baby Shark - Kids' Songs & Stories | 68200000 | 38843229963 | 2865 | 1400000 |
| 18 | ChuChu TV Nursery Rhymes & Kids Songs | 65900000 | 45757850229 | 633 | 1300000 |
| 19 | Shemaroo Filmi Gaane | 65600000 | 28648024439 | 8502 | 764900 |
| 20 | Colors TV | 64600000 | 61510906457 | 112915 | 3600000 |

**Outputs:**

We have performed simple linear regression on above data and got below results:

Equation for Multiple regression: y = -793433.63 + 0.02 (subscribe) +0.000013 (Video Views) + 21.58 (Uploads)



**Takeaways of Multiple Regression Analysis:**

-793433.63

0.02

0.000013

21.58

**Standard error (**

1101318.063

Hight standard error suggest us that **our Multiple regression model** of **YouTube earning as a function of Subscriber, Video Views and Uploads is not the best fit.**

**coefficient of determinant (R2)**

To measure the strength of the relationship we use coefficient of determinant (R2)

R2 = 0.7643

**Which means 76.43% variation YouTube earning is expected by the variance in Subscriber count, Video Views and Uploads and remaining 23.57 is unexpected, due to error.**

Coefficient of determinant does not have a critical value that enable us to draw conclusion however higher the R2 better the model fits data.

If it R2 is 1 then it would be considered as perfect match and if R2 is 0 then it would be considered as no linear relation.

In our case R2 is near to 1 hence we can say that model is better fit for the data taken.

**Test to determine if there is a linear relationship between the YouTube earning and Subscribe, Video Views and Uploads (at 5% significant number):**

**: = = = 0**

**: 0**

Null hypothesis is not true, since none of the independent variables is linearly related to y, and so the model is valid.

**F-statistic**

**F-statistic** of 17 indicates that at least one of the predictor variables has a significant effect on the response variable. This suggests that the overall regression model is statistically significant.

In conclusion, based on the analysis, the model seems to have some validity, as indicated by the overall significance of the regression, hypothesis, the R-squared value, and the significance of certain predictor variables.

**REFERENCE**

Keller, G. (2016). Statistics for Management and Economics (12th ed.). CENGAGE.

Course content of MGT 2148 - Business Analytics and Decision Making

zedstatistics. Learn Statistical Regression in 40 mins! My best video ever. Legit. YouTube <https://www.youtube.com/watch?v=eYTumjgE2IY&ab_channel=zedstatistics>

SergioGarcia, PhD.. statsprof - multiple linear regression in excel. YouTube. <https://www.youtube.com/watch?v=l1AoR6IaeKE&ab_channel=SergioGarcia%2CPhD>

Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). Introduction to the Practice of Statistics (6th ed.).

Howell, D. C. (2013). Statistical Methods for Psychology (7th ed.). CENGAGE.