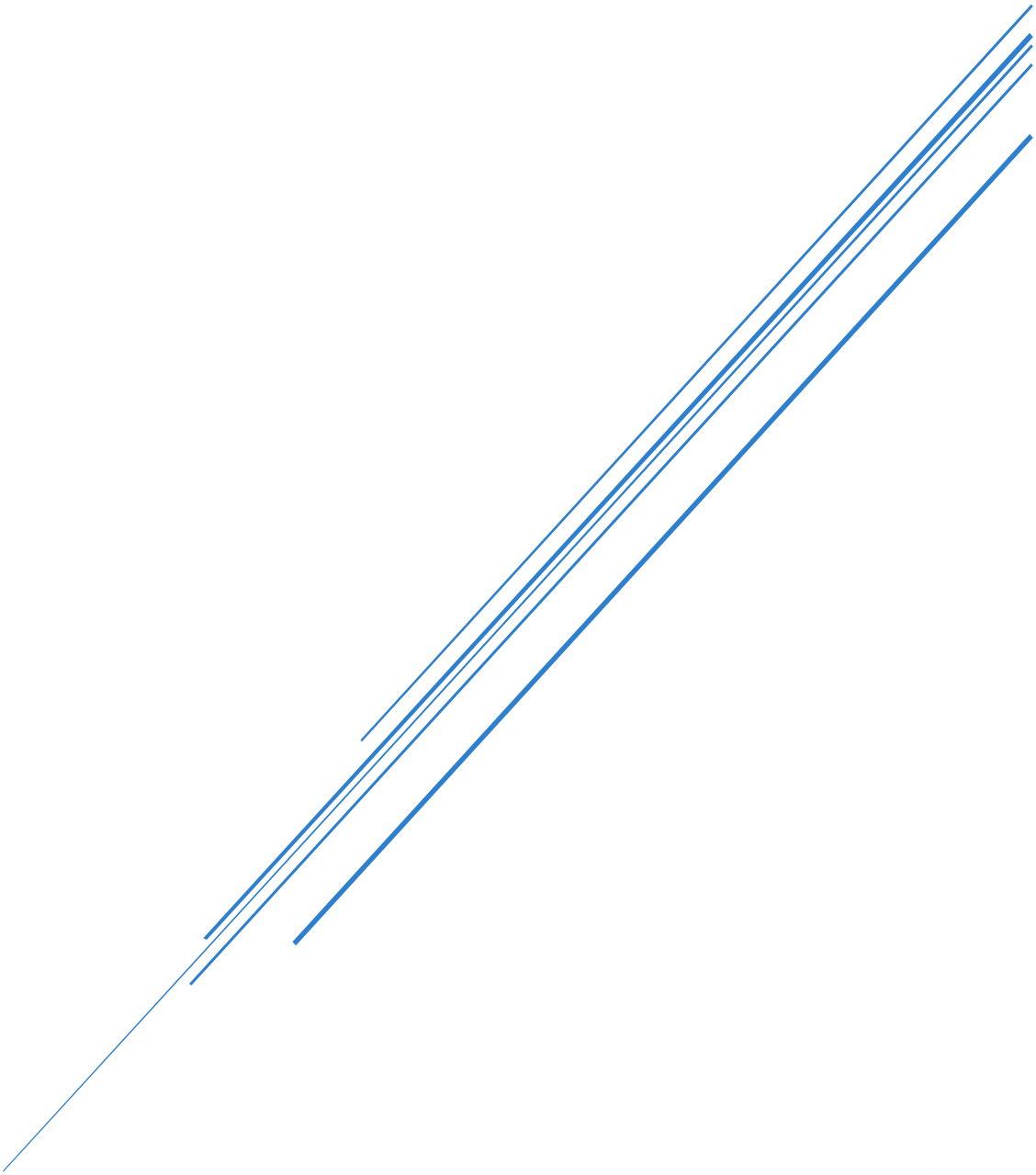
ASSIGNMENT 1

SID: 540771859, 530451954, 530566959



Sydney University
Principals of Data Science

1 540771859 – VEHICLES SALES DATASET

1.1 TOPIC AND RESEARCH QUESTION

Amidst the ongoing increase in the cost of living, it is important to analyse second-hand consumer markets to determine buying behaviour on essential items. This notably extends to the automotive industry, where it is paramount to understand how the second-hand market can shed light on consumer purchasing habits and interests.

To analyse the purchase behaviour of consumers this study focuses on *how distinct factors such as the vehicle makes, market price and vehicle condition can influence the selling price of second-hand vehicles* in *the US market.* Through looking into this market, both the consumer market and automotive producers will be able to gain meaningful insights on the current projections for vehicle purchases as consumers will be able to easily valuate vehicles and gauge purchase price while the automotive industry will be able to determine market trends and consumer preferences in vehicle purchases.

1.2 VEHICLE SALES DATA

1.2.1 Data Provenance and Licence

The 'Vehicle Sales' dataset was sourced from Kaggle, uploaded by Syed Anwar an Artificial Intelligence Engineer at JMM Technologies Limited [1]. According to the author, the file was originally found on a different website (unreferenced) and was uploaded to Kaggle on the 21st of February 2024. Since its upload, the dataset has been viewed 33k and downloaded close to 9500 times with 25 analysis contributions.

The dataset has a usability score of 10/10 by the community and has a Massachusetts Institute of Technology (MIT) licence indicating that it can be copy, modify, merge, publish, distribute, sublicense, and sell copies of the dataset if the copyright licence is attached [2]. Links to both the raw dataset and licence can be found in Appendix 5.1. Though the MIT licence gives users freedom with the dataset, it is worth noting that the data comes with no warranty or liability indicating that analysis of the dataset can lead to poor or an incorrect understanding of the automotive second-hand consumer market.

1.2.2 Data Structure

The raw dataset contains 16 factors ranging from details about the sold vehicles such as the brand and make of the vehicle to the market value of the and selling price. It includes 558837 entries covering the second-hand automotive market from different states in the U.S.A. A data dictionary of the car_sales.csv can be found in Appendix 5.1. Initial observation of the dataset and its structure, indicates that there are a few assumptions and transformations that need to be made to data for modelling.

The 'Make,' 'Model,' 'Trim,' 'Body,' 'State,' 'Color' and 'Interior' all are categorical String variables that can be converted into dummy variables for analysis. As there are only two types of 'Transmission,' the values can be converted from a String to a nominal category as '0' for manual and '1' for automatic. It is key to note that the null values in the dataset represent manual transmissions. According to the contributors on Kaggle, 'Condition' of the vehicle is not given any context within the data collection on how it is determined or what the values mean. A preliminary check of range indicates the values go from 0-50, for this analysis it is assumed that the larger the value, the better the condition of the vehicle is. The last factor that is worth noting is the sale date, where the dataset highlights the sale of cars from December 2014 to July 2015 in 'String' format, which will need to be changed to a 'Date' format.

1.3 DATA CLEANING AND EXPLORATORY ANALYSIS

To start the data cleaning process, the first step was to check the number of missing values within the data frame for each factor. Running a null counts check indicated that the data frame was missing values in all factors with most empty cells relating to the details of the vehicles being sold. As the dataset was originally 558837 datapoints in size, it was decided to drop the rows with missing values as it would have been inaccurate to try and determine

the model/make/trim/body/condition of the vehicles being sold. After dropping the null values, except for transmission where null values indicated that the vehicle has a manual transmission the number of remaining datapoints dropped to 533648. To fix the null values in transmission variable was switched to an integer with '0' representing 'Manual' and '1' representing 'Automatic'.

The next important change that was made was to do with the 'saledate' column of the data frame. As it was the dates were recorded as 'Tue Dec 16 2014 12:30:00 GMT-0800 (PST)', Python was not easily able to switch it from an 'Object' typing to datetime format plus it did not match with the date format given in the 'Year' column. Therefore, to fix this, a function was written which would remove all components of the date so that it became formatted as day/month/year and as seen in Figure 11 in Appendix 5.1 Using this conversion, a new column was then created in the data frame that would calculate the age of the vehicle which could be used for future analysis. When checking the distribution of the Age of vehicles being sold, the data revealed that 174 vehicles had a negative age value, indicating that the data for those rows had inputted incorrectly and hence were removed from the data frame.

The next step was then checking the distribution of other factors within the data frame, starting with 'Selling Price' of the vehicles. An exploratory analysis of the distribution of prices as seen in Figure 1. Indicates many outliers in the dataset as there is a large positive skew which in future cause accuracy and performance issues with machine learning models.

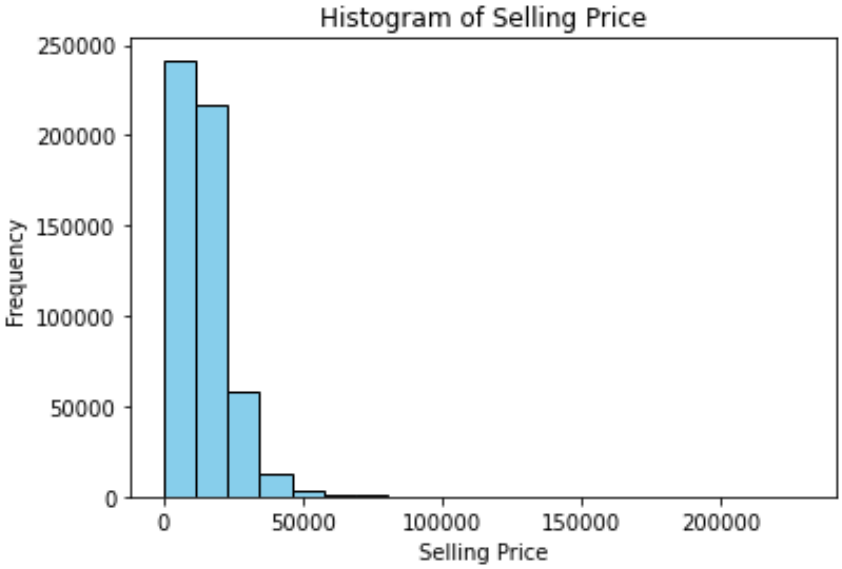


Figure 1. Histogram of Selling Price

Normally to correct the skew to become a normal distribution, outliers would be removed individually to keep as much data references points as possible, but in the context of this study car manufacturers will sell at different prices with high end brands like Lamborghini selling at much higher values. Therefore, an average was taken and two standard deviations on the positive and negative ends were used for the cut off points to create a close to normal distribution within the Selling Price with no outliers as seen in Figure 3.

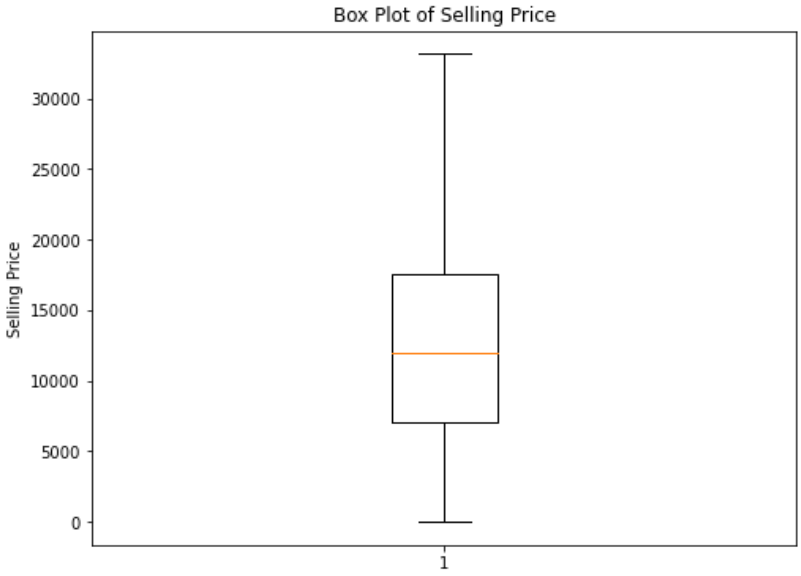


Figure 2. Box Plot of Selling Price

The last step in the data cleaning process and exploratory analysis was to check the distributions of the remaining variables and to drop any columns in the dataset that would no longer be necessary. For instance, the 'vehicle identification number (VIN),' 'saledate' and 'year' were no longer needed as the age would create a more standardised method of comparison between vehicles and the VIN is a unique string with no useful insights for analysis.

Checking the distributions of the remaining factors in the data frame, the remaining factors all are positively skewed except for vehicle 'condition' having a negative skew. The context of the data would support the categorical factors in the dataset to be skewed in their distributions due to the nature of the data, meaning that transformations on the data could result in the ML models identifying trends with unique cases in the dataset.

Though, unlike the categorical factors, the market value for the vehicles (MMR) was expected to be normailsed with the selling price but exploratory analysis through histograms indicated skewness of MMR. Therefore the interquartile method to remove outliers was re-implemented to remove rows in the data with the outliers to normalise the MMR as seen in Figure 3.

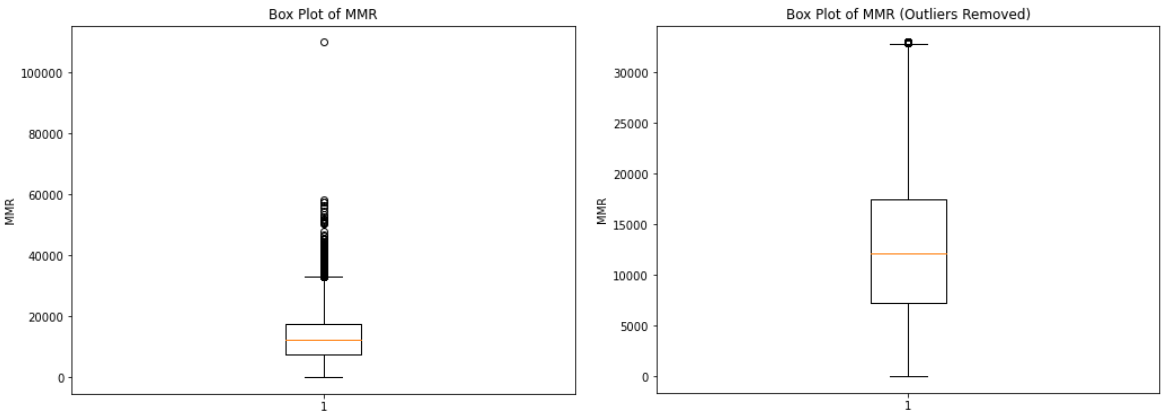


Figure 3. Box Plot of MMR before and after cleaning

2 530451954– ARGENTINA SALARY SURVEY

2.1 TOPIC AND RESEARCH QUESTION

Nowadays, with the rapid development of the internet era, an increasing number of individuals are venturing into the IT industry for employment opportunities. Whether they are IT graduates or professionals considering a career transition, their primary concern revolves around securing high-paying positions. Addressing this inquiry, this study aims to focus on w*hat are the key factors for obtaining high salaries in Argentina's IT sector.*

This study can provide valuable insights for stakeholders, including job seekers and employers, regarding the factors influencing salaries within the IT sector, such as education level, years of experience, job roles, and so on. Job seekers can understand their employment prospects and market preferences, make informed career decisions and plans, and employers can determine market salary levels and provide reasonable and competitive salaries to attract high-quality talents.

2.2 ARGENTINA SALARY SURVEY DATA

2.2.1 Data Provenance and Licence

The 'Argentina Salary Survey' dataset was sourced from Kaggle, and the data originates from OpenQube, a platform dedicated to providing updated information about jobs in the information technology (IT) sector. OpenQube conducts salary surveys within the tech community since 2014, gathering salary-related data from IT professionals. These survey findings are published on the Sysarmy blog. OpenQube selects relevant statistical data from these surveys for public accessibility.

The dataset has the CC BY-NC-SA 4.0 license. According to the license, it is free to share and adapt. However, commercial use of the material is prohibited. If the material is remixed, transformed, or built upon, the resulting contributions must be distributed under the same license as the original. Additionally, no further restrictions may be applied to limit others from exercising the freedoms granted by the license. It's important to note that certain elements of the material may not fall under the license, and the license does not offer warranties or cover other rights such as publicity, privacy, or moral rights.

2.2.2 Data Structure

This dataset contains 5422 rows, and each row has 47 attributes. It covers personal characteristics of employees, such as age and gender, as well as work-related information such as work location, contract type, salary details, years of experience, as well as the technologies and tools used in their roles. Additionally, it involves aspects such as the educational background and guard charges. As this is a survey questionnaire, it also includes survey responses related to salary satisfaction, job-seeking status, among other factors. The data dictionary of the Argentina Salary Survey.csv found in Appendix 5.2.

There are three types of data in this dataset: nominal, ordinal, and ratio. Variables such as Salary, Last Monthly Net Salary, Last Dollar Value, Total Accumulated Adjustment, Years of Experience, Age, etc., are ratio data, which represent amounts, quantities, or years. Income comparison/satisfaction, recommendation, and usage of AI are ordinal data, where higher numbers indicate higher levels. The remaining variables, such as work location and tools usage, are nominal data. Since the data and attributes of this dataset from Argentina are entirely in Spanish, the mentioned attributes use the meanings of each column name for reference.

2.3 DATA CLEANING AND EXPLORATORY ANALYSIS

After downloading this dataset from Kaggle, it can be seen that the name of this csv file is very long. For ease of use, its name has been changed to 'Argentina Salary Survey.csv'. Before starting data cleaning, it is necessary to check this dataset. As this is an Argentine dataset, all data and attributes are in Spanish, so it is necessary to translate the column names and some data into English. First, pandas are used to read the CSV file and check the column names and data types. Use a dictionary to map Spanish column names to English and replace column

names. However, some column names were too long, so dictionary mapping was used again and replaced with shorter English names.

Second, check for missing values in the data. It's noted that there are 17 rows with missing values, most of which exceed 50%, with ‘last_dollar_value’ exhibiting over 80% missing values. Considering that the exchange rate of the US dollar is not directly related to salary levels, ‘last_dollar_value’ is deleted. Similarly, some attributes without impact on salary levels, such as ‘paymentsin_dollars’, ‘Continue_Answering’, ‘income_satisfaction,’ etc., have also been removed. Upon comparison, it's observed that the data of ‘last_monthly_salary_or_gross’ are entirely consistent with salary‘s, so it is removed. Given that all ‘work_countries’ are from Argentina, this column is also deleted.

Next, the five tool-related columns with few missing values are filled using the mode. The high correlation between 'last_monthly_salary_or_net' and 'salary' prompts the utilization of a linear regression model to impute missing values by leveraging the relationships between 'salary' and the target variable in the dataset. In essence, missing values of 'last_monthly_salary_or_net' are predicted using the available data.

Finally, since there are a considerable number of missing values in 'max_studies_level' and 'state', but they still hold analytical value, the decision was made to directly delete the rows with missing values. The processed dataset includes 20 columns and 1927 rows.

Since our research focuses on salary, we conduct exploratory data analysis on salary. It is worth noting that there is a column indicating that salaries are divided into dollarized and non-dollarized, which means that the units of salary data are different and need to be analyzed separately.

The first step is to view the data summary by analyzing the basic statistics of salary, such as mean, median, standard deviation, maximum and minimum values.

The second step is to visualize the data and use a histogram to view the data distribution, as shown in Figure 4.

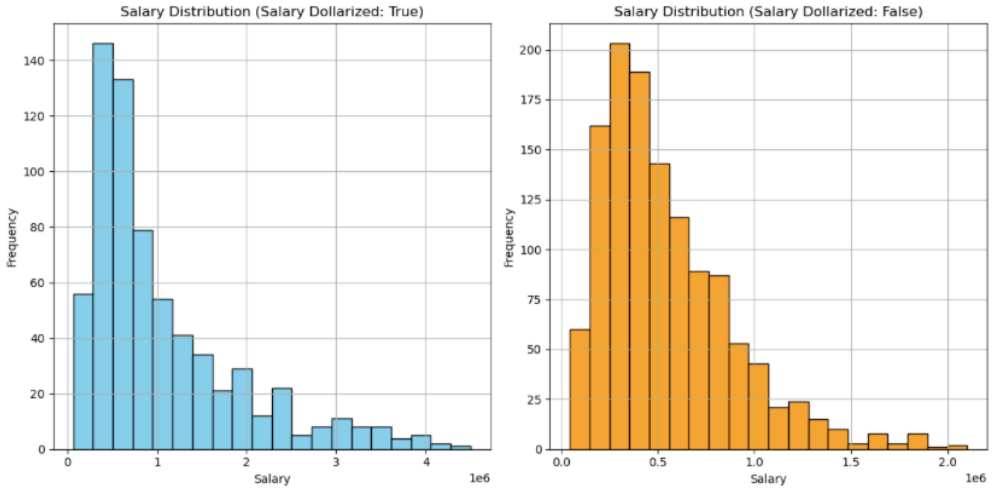


Figure 4. Histograms of Salary Distribution

The data shows a long-tailed distribution, indicating the presence of some outliers. In this case, box plots can be used to identify and remove outliers, making the data more in line with a normal distribution. As shown in Figure 5.

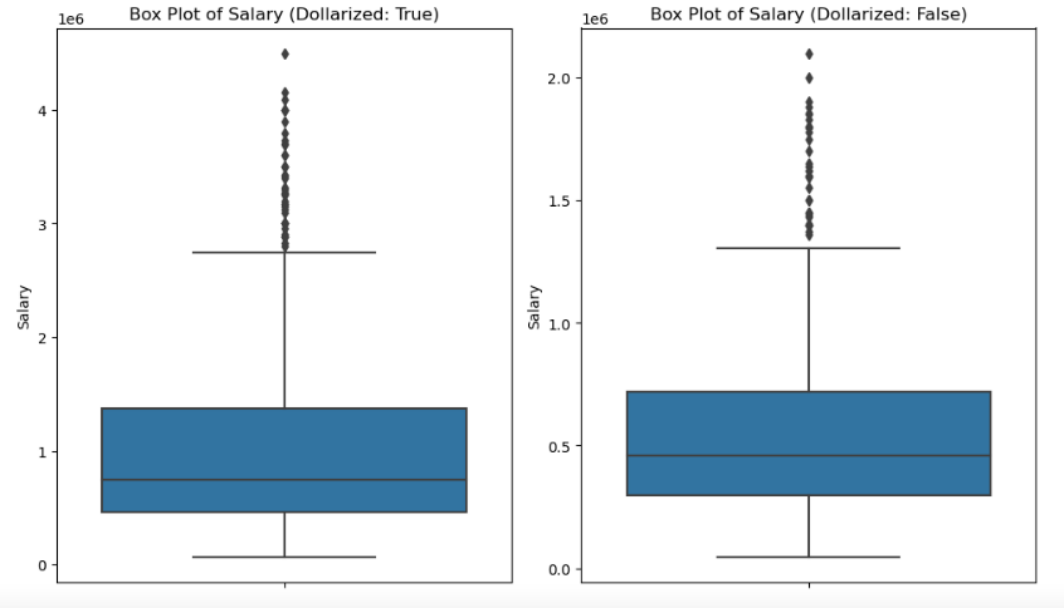


Figure 5. Box plot of Salary's Distribution

This chart shows the presence of outliers outside the box. To clear the outliers, the corresponding quartiles (Q1 and Q3) and interquartile range (IQR) were calculated for the 'salary_dollarized' true and false salaries, respectively. Then, the upper and lower bounds of the outliers were calculated. Next, use Boolean index filtering to select data that is not within the outlier range and remove it from the dataset. Figure 6 shows the salary distribution after handling outliers.

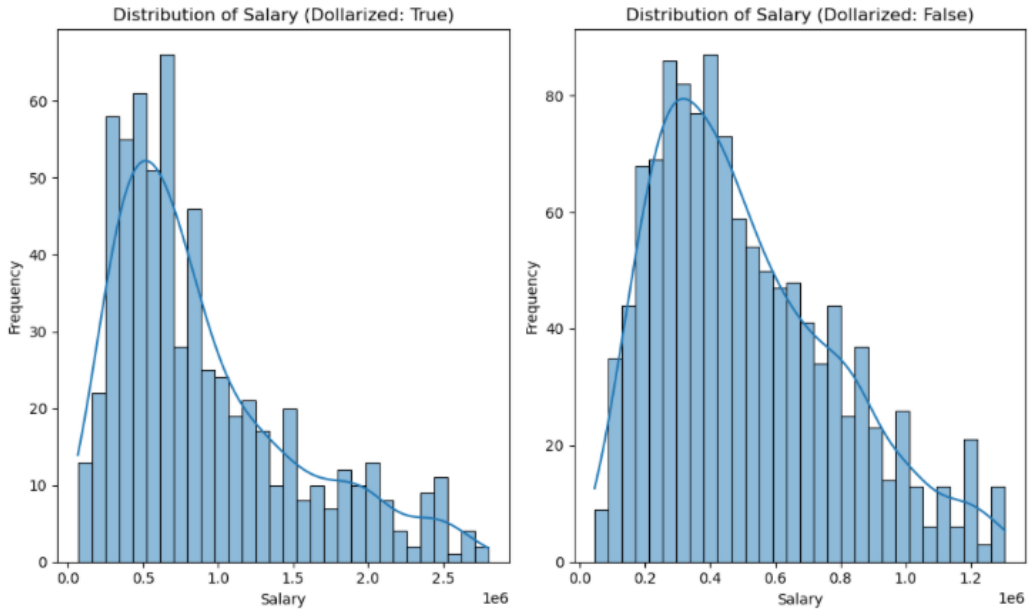


Figure 6. Histograms of Salary Distribution After Processing

The distribution has improved and is closer to a normal distribution, which helps to ensure that subsequent data analysis and modelling processes can be based on more robust data.

3 530566959 – GLOBAL YOUTUBE STATISTICS 2023

3.1 TOPIC AND RESEARCH QUESTION

What are the core elements of a successful YouTube channel? Explore how factors such as category, country, and creation time affect your channel's subscriber count and video views. A successful YouTube channel is not only determined by superficial numbers, but also by the innovation, interactivity, and market positioning of the content. The research question is: What key factors are crucial for a YouTube channel to achieve a high number of subscribers and video views? Is there consistency across different categories of factors and across countries?

3.2 GLOBAL YOUTUBE STATISTICS 2023

3.2.1 Data Provenance and License

The 'Global YouTube Statistics' dataset was sourced from Kaggle on 11 March 2024. It was uploaded by Nidula Elgiriyewithana, a data scientist. According to the author, the file was originally from multiple data sources (unreferenced). Since its upload, the dataset has been viewed 161k and downloaded close to 33.6k times.

The license of the dataset is not specified. It is free to share and adapt. However, commercial use of the material is prohibited. If the material is remixed, transformed, or built upon, the resulting contributions must be distributed as the original. The data comes with no warranty or liability indicating that analysis of the dataset can lead to poor or an incorrect understanding of the factors of a successful YouTube channel.

3.2.2 Data Structure

The raw dataset contains 28 factors. It includes 1000 entries covering the YouTube channels from different countries. A data dictionary of the dataset can be found in Appendix 5.3. The attribute set not only includes basic quantitative indicators, number of subscribers and views, but also extends to more specific dimensions, such as the channel to which the channel belongs.The richness and detail of the data set also provide a unique perspective for exploring internationalization and localization phenomena in the YouTube ecosystem. Comparative analysis of channels in different countries and regions can reveal the complex network of global cultural exchanges and the way local culture is presented and spread on global platforms.

3.3 DATA CLEANING AND EXPLORATORY ANALYSIS

During the data cleaning phase, an exhaustive preliminary review was conducted for missing values in each column. This process was imbued with reverence for the integrity of the data and the accuracy of the analysis. Among the observations, the number of missing values in the country and abbreviation columns is particularly noticeable. In order to maintain the coherence of the data and avoid excessive speculation on unknown information, a unified solution was chosen: these missing values are uniformly marked as "unknown". This practice not only prevents unnecessary guessing about the data set, but also establishes a cornerstone of authenticity and reliability for the analysis.

Next, turn to the cleaning work of the two fields of video views and uploads. Given the central role of these two variables in quantitative analysis, careful filtering of all non-numeric characters ensured that the data in these columns fully conformed to the numeric type requirements. This standardization of data types not only improves the operability of the data, but also lays an solid foundation for future quantitative analysis.

Inconsistent data in the category and country columns were carefully cleaned and normalized. All values that do not meet the predefined standard categories are uniformly marked as 'unknown', thereby eliminating noise introduced by improper classification and ensuring the accuracy of data analysis results. These initiatives effectively solve the problem of data consistency and provide a clear, standardized data set, which creates conditions for in-depth research on a stable and unified basis.

When performing Exploratory Data Analysis (EDA), use the describe() function to obtain basic statistics for numeric columns, which gives a high-level overview of the data set. Visualizing the subscriber column through a

box plot not only reveals the overall trend of the data distribution, but also helps identify potential outliers, as shown in Figure 7.

The resulting boxplots provide insights into the subscriber distribution of YouTube channels. It shows that the main cluster of subscriber numbers is very tight and concentrated in the lower subscriber range, which reveals the remarkable finding that most channels only have a limited subscriber base. At the far end of the graph, outliers are observed indicating that a few channels have huge subscriber numbers. Not only do these data points disrupt convention, they also challenge our conventional understanding of the concept of 'average'.

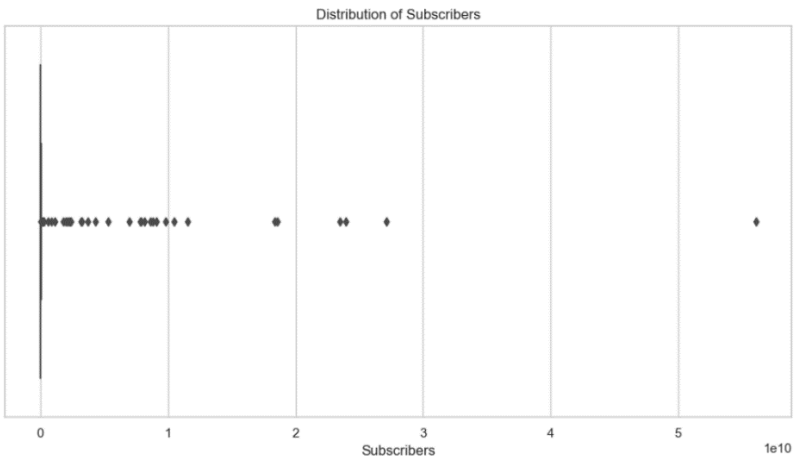


Figure 7. YouTube channel category distribution

3.4 Exploratory Data Analysis (EDA)

Based on the examination of YouTube channel category distribution in Figure 8, it is observed that the number of channels in the entertainment, music, and personal blog categories dominate. Trends seem to firmly indicate a preference among users for entertaining and emotionally resonant content. However, it reflects an issue worthy of attention from a critical perspective: whether the public's attention is too focused on easy entertainment and neglects content in fields such as education, science, and cultural depth. Such a bias may lead to a simplification of the content ecology, thereby inhibiting the diversity and depth of creativity.

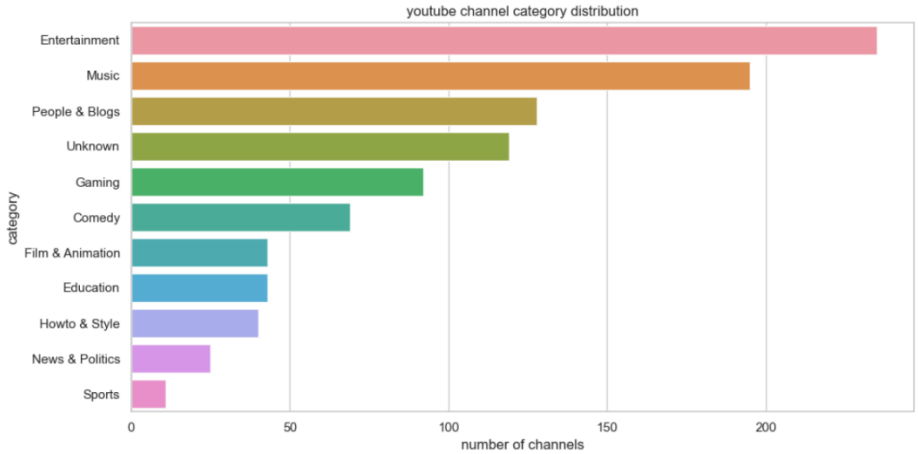


Figure 8. YouTube channel category distribution

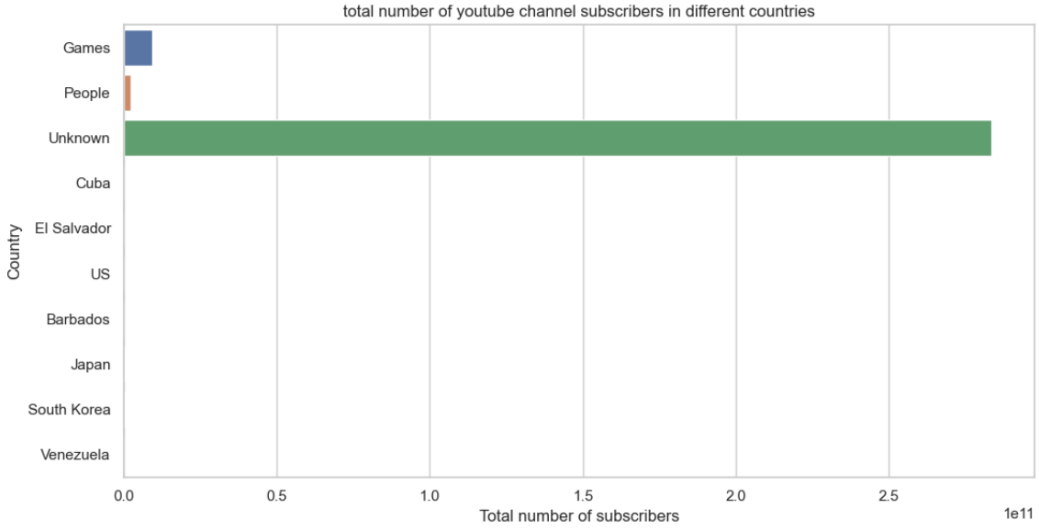


Figure 9. Total number of YouTube channel subscribers in different countries

As shown in Figure 9, the relationship between the year of YouTube channel creation and the average number of subscribers shows that channels established early have a longer time to accumulate audience content and have more subscribers on average. It was believed that the longer the channel history, the greater the success.

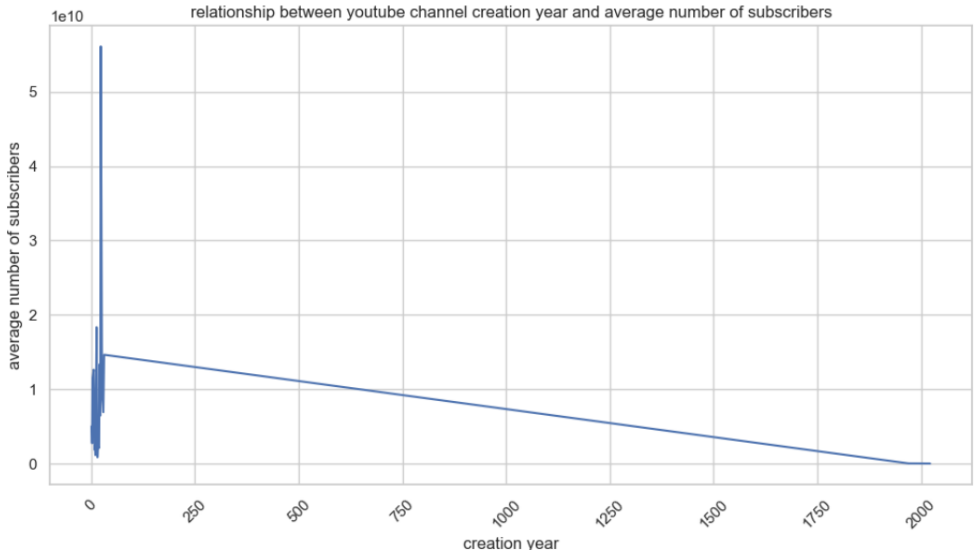


Figure 10. relationship between YouTube channel creation year and average number of subscribers

These figures are essentially a mirror to the preferences of viewers, cultural inclinations, and strides in technology. Habit of leaning on these metrics to gauge a channel's triumph might be misleading, as reality presents a much more intricate picture. For instance, a channel boasting a hefty subscriber count. It's easy to label it a success at first glance. Yet, this perspective might be too shallow. Some channels, despite a modest following, wield significant sway in sectors like education or technology, eclipsing the reach of more mainstream channels. This situation prompts a revaluation of what it truly means to be "successful." To grasp a fuller picture, stepping back and adopting a new viewpoint is crucial. Engaging directly with viewers to harvest genuine reactions to content reveals their true interests—spanning beyond mere entertainment to encompass the calibre and substance of what's being offered. Furthermore, examining which channels have the prowess to continually attract fresh faces while keeping their loyal audience engaged shines a light on their capacity for innovation and adaptation.

SID: 540771859, 530451954, 530566959

4 DISCUSSION AND CONCLUSION

An initial glance each of the datasets each of the raw datasets appeared to be fit for machine learning and future prediction analysis, but through the cleaning process of each of the chosen datasets it was clear that all the datasets were presented with relative positives and negatives when compared with each other which could influence the accuracy of future analysis. These are outlined below:

4.1 VEHICLE SALES DATA

4.1.1 Advantages:

- Volume of Data: After the data cleaning process, the 'vehicle_sales' dataset contains 500,000+ data points with 14 factors.
- Context on Factors: The context on what each factor represents is provided by the original data author with clear naming conventions except for 'condition' which has a range of 0-50. The assumption for this study is that the closer to a value of 50, the better the quality of the vehicle.
- Dependent Factor ('Selling') is normally distributed: After data cleaning, the target variable of 'Selling' has a slight positive skew but can be seen as normally distributed which is a requirement for machine learning.
- Diversity of factors: The data has a large diversity of unique values, particularly in 'model' and 'age' which could give good insights onto how unique cases could influence the selling price of vehicles.

4.1.2 Disadvantages:

- Skewed Independent Factors: Unlike the dependent variable, the independent factors that could be used to predict 'selling' all appear to have a positive skew except for 'condition' with a negative skew are not close to normally distributed. Though this might give greater insights into the small cases of when a rarer vehicle is listed on the market, it can reduce the overall accuracy of the ML models.
- Multi-Collinearity between Factors: Due to the nature of the data, factors such as 'model' and 'make' will be closely linked and be able to predict each other as each car manufacturer is directly linked with the make of the car being produced. This could lead to factors predicting each other rather than the selling price of the vehicle.
- Categorical factors: Due to the large number of categorical factors, applying models can be quite difficult as dependent on the model, it may be appropriate to introduce dummy variables to check for each possible selection method, which could exponentially increase the data used for analysis and complexity of the model.

4.2 ARGENTINA SALARY SURVEY

4.2.1 Advantages

- Large data volume: The dataset contains 5422 rows, covering a considerable number of samples, which is helpful for statistical analysis and modelling.
- Diversity: The dataset contains a wealth of attributes, covering various aspects such as job performance, salary information, personal characteristics, etc., which helps to comprehensively understand the salary situation in the Argentine IT industry.

4.2.2 Disadvantages

- Existence of missing values: Although there are relatively few missing values overall, certain variables such as "pagos_en_dolares" and "cuanto_cobras_por_guard" have many missing values, which is not conducive to analysis.
- Unclear variable naming and unclear meaning: Some variable names are long and vague and need to be converted to clearer and more explicit naming to improve the readability and comprehensibility of the dataset.

- Data redundancy: There are a large number of attributes that are not related to the research, such as' salier_o_seguier_contestando ', indicating whether to continue answering.

4.2.3 EDA

- For EDA, histograms can visually display the distribution of data, effectively demonstrating the concentration, symmetry, and skewness of the data. Box plot can visually identify outliers in data and display statistical information such as median, quartile, and extremum. The display effect of histograms on outliers is not as good as that of box plots, and the display effect is influenced by the amount of data and grouping method. Box plot cannot display specific distribution information.

4.3 YOUTUBE STATISTICS

4.3.1 Advantages

- The diversity is good, with YouTube channels in multiple categories from different countries, the success factors of YouTube channels can be analysed from a global perspective. It has good richness and comprehensive information about the channel, including the number of subscribers, number of video views, categories, countries, and creation time, which are important dimensions for the success of the channel. It is real-time and the data is relatively current, so you can understand the dynamic trends of the YouTube market.

4.3.2 Disadvantages

- Missing values. There are a certain number of missing values in the data set. In the Country key column, it will affect the accuracy of analysis, and the data quality is also a bit poor. Video views and uploads have non-numeric characters, which requires data cleaning and increases the workload. Moreover, the categories are inconsistent. There are some non-standard category garbled characters in the category column. Additional cleaning is required to ensure that the categories are consistent.

4.4 CONCLUSION

Looking at the three datasets described in this report, the positives and negatives outlined of each highlight how they could all be used for analysis with further transformations to the datasets. For instance, to mitigate the missing values and contexts in the Argentina Salary Survey and YouTube statistics would be to decide whether the columns with missing values or number of rows are more important in ML modelling process. Though, when comparing the three sets, the Vehicle_Sales dataset is a more complete dataset that does not carry the same issues as the other two as it has no missing values and clear contexts with the meaning of each factor. The main downside of this dataset is to do with the skew of the independent factors, but with future cleaning and reduction of needed columns dependent on the ML algorithms implemented, these disadvantages can be mitigated for increased accuracy in the models.

Therefore, for the next stage of this project, the group has selected to use the ‘Vehicle_Sales’ dataset to determine how different factors can influence the sale price of vehicles in the second-hand market within the United States of America.

5 APPENDIX

5.1 APPENDIX – 540771859

Link to raw dataset: https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

Link to MIT Licence: https://www.mit.edu/~amini/LICENSE.md

Table 1. Data Dictionary of Raw Vehicle Sales Dataset

Field Name	Type	Data Type	Description	Example
Year	Nominal	Int	Year the car was built.	2015
Make	Nominal	String	The car manufactures.	Kia
Model	Nominal	String	Model of the car.	Sorento
Trim	Nominal	String	The car model type.	LX
Body	Nominal	String	Categorisation of the vehicle based on its design and size.	SUV
Transmission	Nominal	String	The transmission type of the car.	Automatic
Vin (Vehicle Identification Number)	Nominal	String	Unique Vehicle Identifier.	5xyktca69fg566472
State	Nominal	String	State in the U.S.A.	ca
Condition	Ratio	Int	A range from 0-50 listing the condition of the vehicle.	5
Odometer	Interval	Int	Number of miles the vehicle has done.	16639
Color	Nominal	String	The outside colour of the car	White
Interior	Nominal	String	The interior colour of the car.	Black
Seller	Nominal	String	The re-seller of the vehicle.	kia motors america inc
MMR (Manheim Market Report)	Interval	Int	The market value of the vehicle.	20500
Selling Price	Interval	Int	The price the vehicle was sold.	21500
Sale Date	Nominal	String	The date of the sale.	Tue Dec 16 2014 12:30:00 GMT-0800 (PST)

```python
#Function to transform the date to day/month/year
def Date_Transformation(input_string):
    # Splitting string to remove timezone info
    input_string = input_string.split(" GMT")[0]
    datetime_obj = datetime.strptime(input_string, '%a %b %d %Y %H:%M:%S')
    return datetime_obj.strftime('%b %d %Y')

# Creates a column to calculate the age of the vehicles when sold and drops values <0
df['Age'] = df['saledate'].apply(lambda x: int(x[-4:]))
df['Age'] = df['Age'] - df['year']
print(df['Age'].value_counts())
df = df[df['Age'] >= 0]
print(df['Age'].value_counts())
```

Figure 11 – Date Transformation and Age factor implementation.

Table 2. Data Dictionary of Cleaned Vehicle Sales Dataset

Field Name	Type	Data Type	Description	Example
Make	Nominal	Object	The car manufactures.	Kia
Model	Nominal	Object	Model of the car.	Sorento
Trim	Nominal	Object	The car model type.	LX
Body	Nominal	Object	Categorisation of the vehicle based on its design and size.	SUV
Transmission	Nominal	Int	The transmission type of the car.	1
State	Nominal	Object	State in the U.S.A.	ca
Condition	Ratio	Int	A range from 0-50 listing the condition of the vehicle.	5
Odometer	Interval	Float	Number of miles the vehicle has done.	16639.0
Color	Nominal	Object	The outside colour of the car	White
Interior	Nominal	Object	The interior colour of the car.	Black
Seller	Nominal	Object	The re-seller of the vehicle.	kia motors america inc
MMR (Manheim Market Report)	Interval	Float	The market value of the vehicle.	20500
Selling Price	Interval	Float	The price the vehicle was sold.	21500
Age	Nominal	Int	Age of the Vehicle	2

5.2 APPENDIX – 530451954

Link to raw dataset: https://www.kaggle.com/datasets/aletbm/argentina-salary-survey-sysarmy

Link to MIT Licence: CC BY-NC-SA 4.0

Table 3. Data Dictionary of Raw dataset

Field Name	Type	Data Type	Description	Example
estoy_trabajando_en	Nominal	String	Current Country	Argentina
donde_estas_trabajando	Nominal	String	Current Workplace	Ciudad Autónoma de Buenos Aires
dedicacion	Nominal	String	Work Dedication	Full-Time
tipo_de_contrato	Nominal	String	Type of Contract	Staff (planta permanente)
ultimo_salario_mensual_o_retiro_bruto_en_tu_moneda_local	Ratio	Float	Last Monthly Gross Salary or Withdrawal in Local Currency	660000
ultimo_salario_mensual_o_retiro_neto_en_tu_moneda_local	Ratio	Float	Last Monthly Net Salary or Withdrawal in Local Currency	380000
pagos_en_dolares	Nominal	String	Payments in Dollars	Mi sueldo está dolarizado (pero cobro en moneda local)
si_tu_sueldo_esta_dolarizado_cual_fue_el_ultimo_valor_del_dolar_que_tomaron	Ratio	String	Last Dollar Value Considered for Dollarized Salary	231
recibis_algun_tipo_de_bono	Nominal	String	Receipt of Any Bonuses	De uno a tres sueldos
a_que_esta_atado_el_bono	Nominal	String	Bonus Attachment	Mix de las anteriores
tuviste_actualizaciones_de_tus_ingresos_laborales_durante_2023	Nominal	String	Income Updates During 2023	No
de_que_fue_el_ajuste_total_acumulado	Ratio	Float	Total Accumulated Adjustment	0
en_que_mes_fue_el_ultimo_ajuste	Nominal	String	Month of Last Adjustment	No tuve

como_consideras_que_estan_tus_ingresos_laborales_comparados_con_el_semestre_anterior	Ordinal	Int	Comparison of Current Labor Income with the Previous Semester	2
contas_con_beneficios_adicionales	Nominal	String	Additional Benefits	Abono de Internet, Capacitaciones y/o cursos, Descuento en gimnasios / Clases de gimnasia online, Clases de idiomas, Descuentos varios (Clarín 365, Club La Nación, etc), Horarios flexibles, Lactario en la oficina, Licencia por nacimiento extendida, Guardería o pago para cubrir costo de guardería, Stock options / RSUs, Vacaciones flexibles (adicionales a las reglamentarias)
que_tan_conforme_estas_con_tus_ingresos_laborales	Ordinal	Int	Satisfaction with Labor Income	2
estas_buscando_trabajo	Nominal	String	Job Search Status	No, pero escucho ofertas.
trabajo_de	Nominal	String	Desired Job Field	Manager / Director
anos_de_experiencia	Ratio	Int	Years of Experience	10
antiguedad_en_la_empresa_actual	Ratio	Int	Seniority in Current Company	7
tiempo_en_el_puesto_actual	Ratio	Int	Time in Current Position	7
cuantas_personas_a_cargo_tenes	Ratio	Int	Number of Subordinates	20
plataformas_que_utilizas_en_tu_puesto_actual	Nominal	String	Platforms Used in Current Position	Ninguna de las anteriores

lenguajes_de_programacion_o_tecnologias_que_utilices_en_tu_puesto_actual	Nominal	String	Programming Languages or Technologies Used in Current Position	Ninguno de los anteriores
frameworksherramientas_y_librerias_que_utilices_en_tu_puesto_actual	Nominal	String	Frameworks, Tools, and Libraries Used in Current Position	Ninguno de los anteriores
bases_de_datos	Nominal	String	Databases Utilized	Ninguna de las anteriores
qa_testing	Nominal	String	QA Testing	Ninguna de las anteriores
cantidad_de_personas_en_tu_organizacion	Nominal	String	Number of People in Organization	De 5001 a 10000 personas
modalidad_de_trabajo	Nominal	String	Work Modality	100% remoto
si_trabajas_bajo_un_esquema_hibrido_cuantos_dias_a_la_semana_vas_a_la_oficina	Ratio	Int	Frequency of Office Attendance in Hybrid Work Setup	0
la_recomendas_como_un_buen_lugar_para_trabajar	Ordinal	Int	Recommendation as a Good Place to Work	7
que_tanto_estas_usando_copilotchatgpt_u_otras_herramientas_de_ia_para_tu_trabajo	Ordinal	Int	Usage of Copilot/ChatGPT or Other AI Tools for Work	1
salir_o_seguir_contestando	Nominal	String	Continue or Exit Questionnaire	Responder sobre guardias
maximo_nivel_de_estudios	Nominal	String	Highest Level of Education	Universitario
estado	Nominal	String	Location	Incompleto
carrera	Nominal	String	Field of Study	Licenciatura en Administración de Empresas
institucion_educativa	Nominal	String	Educational Institution	UBA - Universidad de Buenos Aires
salir_o_seguir_contestando_	Nominal	String	Continue or Exit Shifts Section	Terminar encuesta

sobre_las_guardias				
tenes_guardias	Nominal	String	Shifts Availability	No
cuanto_cobras_por_guardia	Ratio	Float	Rate Per Shift	0
aclara_el_numero_que_ingresaste_en_el_campo_anterior	Nominal	String	Clarification of Previous Field Entry	Porcentaje de mi sueldo bruto
salir_o_seguir_contestando_sobre_estudios	Nominal	String	Continue or Exit Education Section	Responder sobre mis estudios
tengo_edad	Ratio	Int	Age	45
me_identifico_genero	Nominal	String	Gender Identification	Hombre Cis
sueldo_dolarizado	Nominal	bool	Dollarized Salary Status	TRUE
seniority	Nominal	String	Seniority Level	Senior
_sal	Ratio	Float	Salary	660000

5.3 APPENDIX – 530566959

Link to raw data: https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023/data
Link to license: This data set has no specific license.

Field Name	Type	Data Type	Description	Example
rank	Quantitative	Integer	The rank of the YouTube channel based on the number of subscribers.	1
Youtuber	Qualitative	String	The name of the YouTube channel.	T-Series
subscribers	Quantitative	Integer	The number of subscribers to the channel.	245000000
video views	Quantitative	String	Total views across all videos on the channel. (May require conversion to numeric type)	2.28E+11
category	Qualitative	String	The category or niche of the channel.	Music
Title	Qualitative	String	The title of the YouTube channel.	T-Series
uploads	Quantitative	String	Total number of videos uploaded on the channel. (May require conversion to numeric type)	20082
Country	Qualitative	String	The country where the YouTube channel originates.	India
Abbreviation	Qualitative	String	Abbreviation of the country.	IN
channel_type	Qualitative	String	The type of the YouTube channel (e.g., individual, brand).	Music
video_views_rank	Quantitative	Integer	Ranking of the channel based on total video views.	-
country_rank	Quantitative	Integer	Ranking of the channel based on the number of subscribers within its country.	-
channel_type_rank	Quantitative	Integer	Ranking of the channel based on its type (individual or brand).	-

video_views_for_the_last_30_days	Quantitative	String	Total video views in the last 30 days. (May require conversion to numeric type)	-
lowest_monthly_earnings	Quantitative	String	Lowest estimated monthly earnings from the channel. (May require conversion to numeric type)	-
highest_monthly_earnings	Quantitative	String	Highest estimated monthly earnings from the channel. (May require conversion to numeric type)	-
lowest_yearly_earnings	Quantitative	String	Lowest estimated yearly earnings from the channel. (May require conversion to numeric type)	-
highest_yearly_earnings	Quantitative	String	Highest estimated yearly earnings from the channel. (May require conversion to numeric type)	-
subscribers_for_last_30_days	Quantitative	Integer	Number of new subscribers gained in the last 30 days.	2000000
created_year	Quantitative	Integer	Year when the YouTube channel was created.	2006
created_month	Qualitative	String	Month when the YouTube channel was created.	Mar
created_date	Quantitative	Float	Exact date of the YouTube channel's creation.	13.0
Gross tertiary education enrollment (%)	Quantitative	Float	Percentage of the population enrolled in tertiary education in the country.	28.1
Population	Quantitative	Float	Total population of the country.	1.366418e+09
Unemployment rate	Quantitative	Float	Unemployment rate in the country.	5.36
Urban_population	Quantitative	Float	Percentage of the population living in urban areas.	471031528.0
Latitude	Quantitative	Float	Latitude coordinate of the country's location.	20.593684
Longitude	Quantitative	Float	Longitude coordinate of the country's location.	78.962880

SID: 540771859, 530451954, 530566959

6 REFERENCES

[1] S. ANWAR, “Vehicle Sales Data,” Kaggle, Peshawar, Khyber Pakhtunkhwa, Pakistan, 2024.

[2] K. Golubic, “What is MIT License?,” Memgraph, 05 June 2023. [Online]. Available: https://memgraph.com/blog/what-is-mit-license. [Accessed 10 March 2024].