Data Modeling Project 1

Proposal for Data Modeling

1. Jiwon Shin
2. Haochong (Harry) Xia
3. Xueqing (Annie) Wu
4. Rafael Dávila Bugarín

Data Set 1

Data Set

The source of this dataset is Kaggle, and the data can be found at the following link.

Data Set Description

A comprehensive dataset with information on YouTube's top creators, including subscribers, views, upload frequency, country of origin, and earnings, is presented in this dataset.

Research questions

1. How does the number of subscribers, views, uploads, and region factors have an impact on the highest monthly earnings for YouTube channels?
    a. Outcome variable: the highest monthly earnings for each YouTube channel
    b. Model to use: multiple linear regression

2. How likely the Youtube channel is going to earn more than $100K given the views, subscribers, region factors?[1]

    a. Outcome variable: Whether the Youtube channel can have more than $100K earning for the highest monthly earnings (the likelihood)
    b. Model to use: logistic regression

```
 Attaching core tidyverse packages              tidyverse 2.0.0
dplyr     1.1.2      readr     2.1.4
forcats   1.0.0      stringr   1.5.0
ggplot2   3.4.3      tibble    3.2.1
lubridate 1.9.2      tidyr     1.3.0
purrr     1.0.2
 Conflicts                        tidyverse_conflicts()
dplyr::filter() masks stats::filter()
dplyr::lag()    masks stats::lag()
Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
Rows: 995
Columns: 28
$ rank                                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,…
$ Youtuber                            <chr> "T-Series", "YouTube Movies", …
$ subscribers                         <int> 245000000, 170000000, 16600000…
$ video.views                         <dbl> 228000000000, 0, 28368841870, …
$ category                            <chr> "Music", "Film & Animation", "…
$ Title                               <chr> "T-Series", "youtubemovies", "…
$ uploads                             <int> 20082, 1, 741, 966, 116536, 0,…
$ Country                             <chr> "India", "United States", "Uni…
$ Abbreviation                        <chr> "IN", "US", "US", "US", "IN", …
$ channel_type                        <chr> "Music", "Games", "Entertainme…
$ video_views_rank                    <int> 1, 4055159, 48, 2, 3, 4057944,…
$ country_rank                        <dbl> 1, 7670, 1, 2, 2, NaN, 3, 1, 5…
$ channel_type_rank                   <dbl> 1, 7423, 1, 1, 2, NaN, 3, 4, 2…
$ video_views_for_the_last_30_days    <dbl> 2258000000, 12, 1348000000, 19…
$ lowest_monthly_earnings             <dbl> 564600, 0, 337000, 493800, 455…
$ highest_monthly_earnings            <dbl> 9.000e+06, 5.000e-02, 5.400e+0…
$ lowest_yearly_earnings              <dbl> 6.800e+06, 4.000e-02, 4.000e+0…
$ highest_yearly_earnings             <dbl> 1.084e+08, 5.800e-01, 6.470e+0…
$ subscribers_for_last_30_days        <dbl> 2000000, NaN, 8000000, 1000000…
```

[1] In the first Appendix 1 you can find the plot of the Montly revenue distribution in USD. These are the cuts analyzed: 839 out of 995 earns < $1m, 339 out of 995 earns <$100K, 192 out of 995 earns <$10K, 170 out of 995 earns <$1k, 89 out of 995 earns $0)

```
$ created_year                                <dbl> 2006, 2006, 2012, 2006, 2006, …
$ created_month                               <chr> "Mar", "Mar", "Feb", "Sep", "S…
$ created_date                                <dbl> 13, 5, 20, 1, 20, 24, 12, 29, …
$ Gross.tertiary.education.enrollment.... <dbl> 28.1, 88.2, 88.2, 88.2, 28.1, …
$ Population                                  <dbl> 1366417754, 328239523, 3282395…
$ Unemployment.rate                           <dbl> 5.36, 14.70, 14.70, 14.70, 5.3…
$ Urban_population                            <dbl> 471031528, 270663028, 27066302…
$ Latitude                                    <dbl> 20.59368, 37.09024, 37.09024, …
$ Longitude                                   <dbl> 78.962880, -95.712891, -95.712…
```

Data Set 2

The Data Source comes from Los Angeles Police Department (LAPD) and it is available in the following link.

Data Set Description

This dataset represents recorded crime incidents in the City of Los Angeles dating from 2020. The source of this data is transcriptions from original crime reports, which are originally documented on paper. Therefore, it is important to acknowledge that there may exist some inaccuracies or errors within the dataset. In certain cases, location information is missing and indicated as (0°, 0°). To safeguard individual privacy, address details are truncated to the closest hundred block. The accuracy of this data is contingent on the accuracy of the information stored in the database.

Research questions

1. Does crime really happened more often after 7pm? How race, age, gender and location affects. The specific case of Los Angeles City, Cal. (2022)
2. What variables determine if a crime will be commited with or without violence? The specific case of Los Angeles City, Cal. (2022) (The team is still deciding if the types are going to merge to create a dichotomous variable or several applying an multinomial logit.)

Glimse of the data

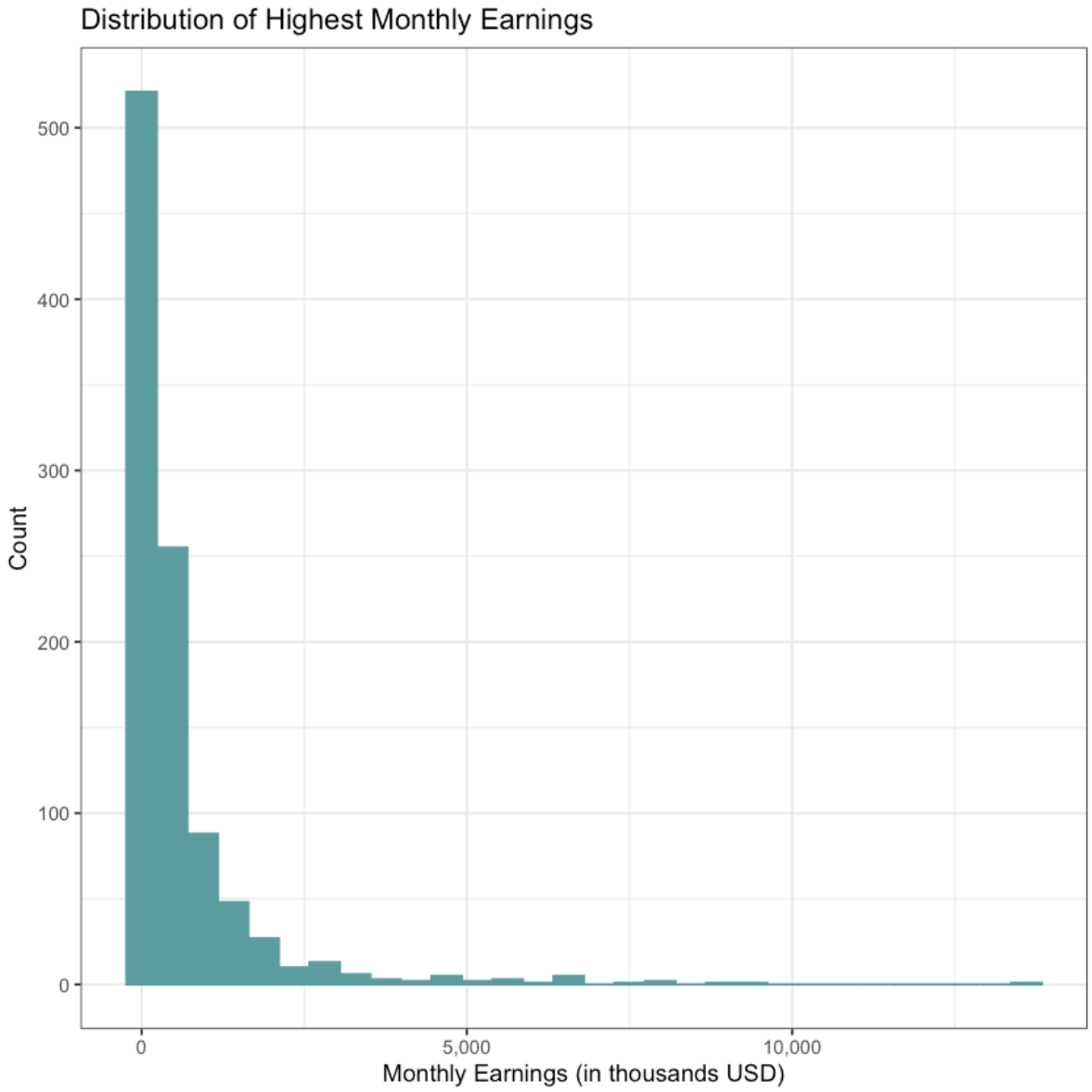
```
Attaching package: 'jsonlite'
```

```
The following object is masked from 'package:purrr':
```

```
    flatten
```

```
Rows: 1,000
Columns: 26
$ dr_no          <chr> "010304468", "190101086", "200110444", "191501505", "19…
$ date_rptd      <chr> "2020-01-08T00:00:00.000", "2020-01-02T00:00:00.000", "…
$ date_occ       <chr> "2020-01-08T00:00:00.000", "2020-01-01T00:00:00.000", "…
$ time_occ       <chr> "2230", "0330", "1200", "1730", "0415", "0030", "1315",…
$ area           <chr> "03", "01", "01", "15", "19", "01", "01", "01", "01", "…
$ area_name      <chr> "Southwest", "Central", "Central", "N Hollywood", "Miss…
$ rpt_dist_no    <chr> "0377", "0163", "0155", "1543", "1998", "0163", "0161",…
$ part_1_2       <chr> "2", "2", "2", "2", "2", "1", "1", "2", "1", "1", "1", …
$ crm_cd         <chr> "624", "624", "845", "745", "740", "121", "442", "946",…
$ crm_cd_desc    <chr> "BATTERY - SIMPLE ASSAULT", "BATTERY - SIMPLE ASSAULT",…
$ mocodes        <chr> "0444 0913", "0416 1822 1414", "1501", "0329 1402", "03…
$ vict_age       <chr> "36", "25", "0", "76", "31", "25", "23", "0", "23", "0"…
$ vict_sex       <chr> "F", "M", "X", "F", "X", "F", "M", "X", "M", "X", "M", …
$ vict_descent   <chr> "B", "H", "X", "W", "X", "H", "H", "X", "B", "X", "A", …
$ premis_cd      <chr> "501", "102", "726", "502", "409", "735", "404", "726",…
$ premis_desc    <chr> "SINGLE FAMILY DWELLING", "SIDEWALK", "POLICE FACILITY"…
$ weapon_used_cd <chr> "400", "500", NA, NA, NA, "500", NA, NA, NA, NA, "306",…
$ weapon_desc    <chr> "STRONG-ARM (HANDS, FIST, FEET OR BODILY FORCE)", "UNKN…
$ status         <chr> "AO", "IC", "AA", "IC", "IC", "IC", "IC", "IC", "IC", "…
$ status_desc    <chr> "Adult Other", "Invest Cont", "Adult Arrest", "Invest C…
$ crm_cd_1       <chr> "624", "624", "845", "745", "740", "121", "442", "946",…
$ location       <chr> "1100 W  39TH                          PL", "700 S  HILL…
$ lat            <chr> "34.0141", "34.0459", "34.0448", "34.1685", "34.2198", …
$ lon            <chr> "-118.2978", "-118.2545", "-118.2474", "-118.4019", "-1…
$ crm_cd_2       <chr> NA, NA, NA, "998", NA, "998", "998", "998", "998", NA, …
$ cross_street   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "OLIVE", NA, NA…
```

Appendix 1


Distribution of Highest Monthly Earnings
Count
500
400
300
200
100
0
0
5,000
10,000
Monthly Earnings (in thousands USD)