Data types

1) Provide a URL to the dataset.

I downloaded my dataset from https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv (https://www.kaggle.com/tmdb/tmdb-movie-metadata#tmdb_5000_movies.csv)

2) Explain why you chose this dataset.

I enjoy watching movies and this was one of the first data sets that caught my eye

3) What are the entities in this dataset? How many are there?

The entities in this dataset are top movies from TMDb. There are 4803 entities

4) How many attributes are there in this dataset?

There are 20 attributes

5) What is the datatype of each attribute (categorical -ordered or unordered-, numeric -discrete or continuous-, datetime, geolocation, other)? Write a short sentence stating how you determined the type of each attribute. Do this for at least 5 attributes, if your dataset contains more than 10 attributes, choose 10 of them to describe.

Num	Name	Type	Description
1	`original_title`	other - title	Title of the movie
2	`tagline`	other - tagline	Tagline of the movie
3	`budget`	numeric continuous	total budget of the movie
4	`status`	categorical unordered	Can take value from finite set of possible statuses
5	`vote_count`	numeric continuous	Total count of votes
6	`vote_average`	numeric discrete	Average vote on a scale of 1-10
7	`release_date`	datetime	Specifies date of release
8	`revenue`	numeric continous	Total revenue earned

Num	Name	Type	Description
9	`runtime`	other - address	Stree address if incident
10	`original_language`	categorical unordered	Can take value from finite set of possible languages

6) Write R code that loads the dataset using function `read_csv`. Were you able to load the data successfully? If no, why not?

```
library(tidyverse)
#When i first used the url provided above, I was receiving a parsing error due to some sort of f
ormat error. However, once I uploaded it from my local machine it worked perfectly
movies <- read_csv('tmdb_5000_movies.csv')
```

Wrangling

1. My pipeline computes the average budget by original language (ignores budgets <=0)

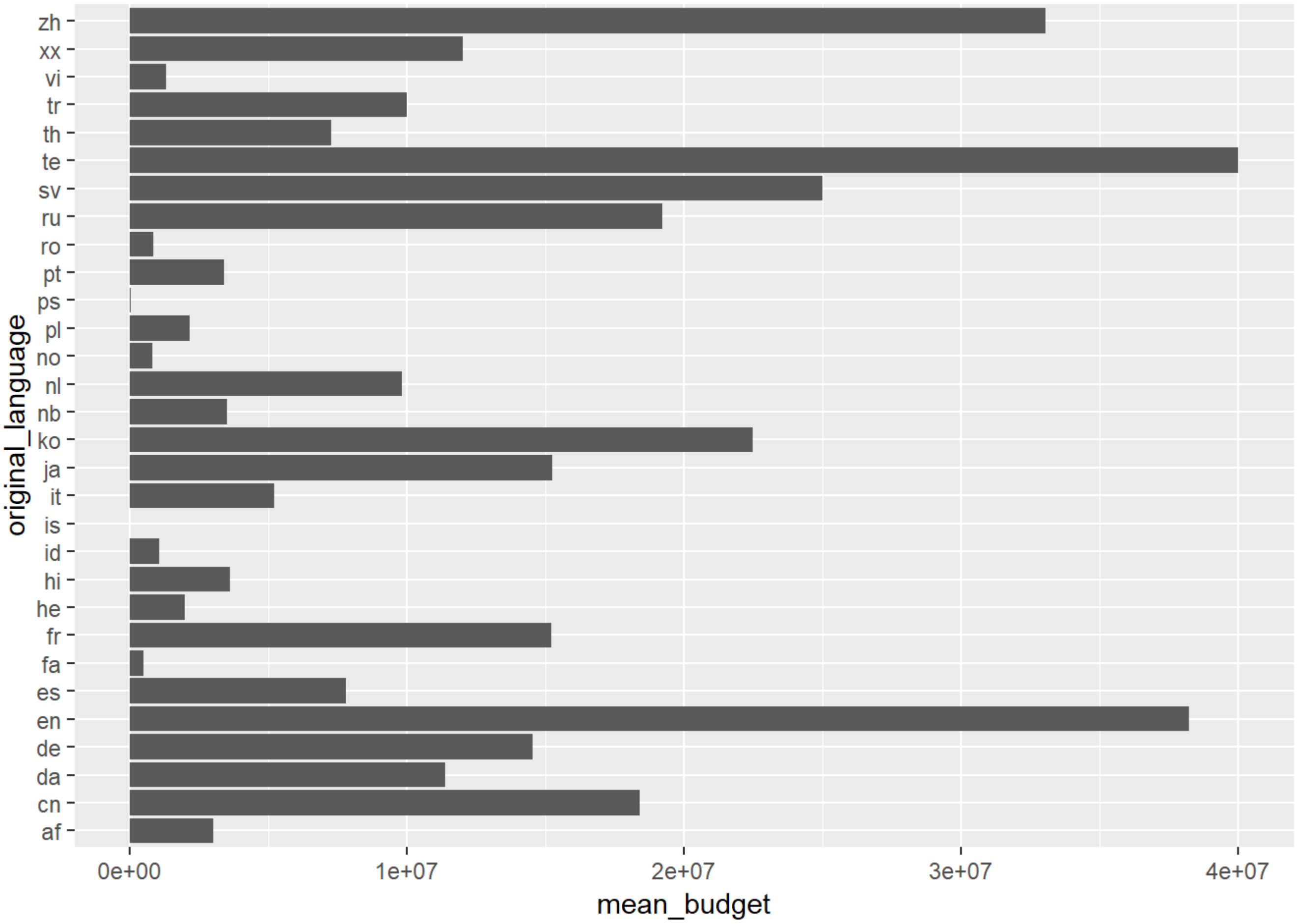
```
mean_budgets <- movies %>%
  filter(budget > 0) %>%
  select(original_language, budget) %>%
  group_by(original_language) %>%
  summarize(mean_budget=mean(budget)) %>%
  arrange(mean_budget)
mean_budgets
```

```
## # A tibble: 30 x 2
##    original_language mean_budget
##    <chr>                   <dbl>
##  1 is                         10
##  2 ps                      46000
##  3 fa                     490000
##  4 no                     800000
##  5 ro                     852510
##  6 id                    1050000
##  7 vi                    1300000
##  8 he                    2000000
##  9 pl                    2159280
## 10 af                    3000000
## # ... with 20 more rows
```

Plotting

1. This barplot shows the average budget per original_language (ignoring budgets <= 0)

```
mean_budgets %>%
  ggplot(aes(x=original_language, y=mean_budget)) +
    geom_bar(stat="identity") +
    coord_flip()
```


zh
xx
vi
tr
th
te
sv
ru
ro
pt
ps
pl
no
nl
nb
ko
ja
it
is
id
hi
he
fr
fa
es
en
de
da
cn
af
original_language
0e+00
1e+07
2e+07
3e+07
4e+07
mean_budget