

```
1  *** I cleaned my data manually in Excel.
2  *** I created my dataset manually in Excel.
3  *****I inputted critics score via Times of India's aggregate
   critics score:
   https://timesofindia.indiatimes.com/entertainment/hindi/movie-review
   S
4  *****I used IMDB for Genre, Cast information, and user popular
   reviews:
   https://www.imdb.com/search/title/?title_type=feature&primary_language=hi&sort=num_votes,desc
5  *****I used Bollywood Hungama for Star Power/Presence Indicator
   Variables: https://www.bollywoodhungama.com/celebrities/top-100/
6  *** import data from Excel
7  cd "/Users/sidravi/Desktop"
8  import excel "/Users/sidravi/Desktop/Hindi Films 2009-2019
   Final.xlsx", sheet("Data") firstrow clear
9  save thesis_data_final.dta, replace
10
11  *** CLEAN DATA IN STATA
12  ***** drop blank columns
13  drop 0 P Q AB AC
14  ***** drop extraneous observations
15  drop in 238
16  drop in 237
17  ***** rename production budget variable*
18  rename movie_financial_summary_product movie_production_budget
19  ***** just in case, I destringed my movie run time, sequel,
   production budget, US Revenue, Indian Revenue, Cast, Ratio,
   Popular Review, star presence and release year, variables
20  destring movie_running_time, replace
21  destring movie_sequel, replace
22  destring movie_production_budget, replace
23  destring movie_US_Revenue, replace
24  destring movie_Indian_Revenue, replace
25  destring movie_women, replace
26  destring movie_men, replace
27  destring movie_ratio, replace
28  destring popular_review, replace
29  destring star_power, replace
30  destring star_power_specific, replace
31  destring star_male, replace
32  destring star_female, replace
33  destring movie_production_year, replace
34  ***** I drop any movies produced/released in 2008*
35  list if movie_production_year == 2008
36  drop if movie_production_year == 2008
37  **** I replace any star power variables with 0 if not in top 25 of
   star rankings list.
```

```
39  replace star_female = 0 if star_female ==.
40
41  ***I rename variables for easier interpretation.
42  rename movie_production_year movie_release_year
43  rename movie_ratio gender_ratio
44  save thesis_data_final, replace
45
46  *** Some movies are listed as "multiple genres" or as a genre
    outside of the five main categories (Action, Comedy, Drama,
    Romantic Comedy, Thriller/Suspense) which interferes with my
    analysis. I went to the IMDb pages of those films and determined
    appropriate genre and then replaced the genre variable value.
47  replace movie_genre = "Thriller/Suspense" in 12
48  replace movie_genre = "Comedy" in 83
49  replace movie_genre = "Comedy" in 102
50  replace movie_genre = "Action" in 106
51  replace movie_genre = "Drama" in 177
52  replace movie_genre = "Comedy" in 182
53  replace movie_genre = "Action" in 222
54  replace movie_genre = "Drama" in 127
55  replace movie_genre = "Comedy" in 158
56  replace movie_genre = "Romantic Comedy" in 164
57  replace movie_genre = "Thriller/Suspense" in 2
58  replace movie_genre = "Comedy" in 26
59  replace movie_genre = "Romantic Comedy" in 38
60
61  tabulate movie_month
62  encode (movie_genre), gen(movie_genre_Num)
63  ***Clean Month Variable, accounting for typos in original Excel
    Dataset*
64  list if movie_month == "Jun"
65  replace movie_month = "Jun" in 6
66  replace movie_month = "Jun" in 167
67  replace movie_month = "Jun" in 77
68  list if movie_month == "Sept"
69  replace movie_month = "Sept" in 110
70  replace movie_month = "Sept" in 161
71  tabulate movie_month
72  *Destring Month Variable, create numerical variable*
73  encode (movie_month), gen(movie_month_Num)
74
75  *** CREATE GENDER STAR PRESENCE VARIABLE
76  *Create "No Star" variable if no male star or female star*
77  gen startype = 0 if (star_male == 0) & (star_female == 0)
78  *Create "Male Star" variable if no female star
79  replace startype = 1 if star_male > 0 & (missing(star_female) |
    star_female == 0)
80  *Create "Female Star" variable if no male star
```

```
81  replace startype = 2 if star_female > 0 & (missing(star_male) |
    star_male == 0)
82  *Create "Both Star" variable if both male and female star is present
83  replace startype = 3 if star_male > 0 & star_female > 0
84  save thesis_data_final, replace
85
86  ***Take Logarithm of Production Budget, US Revenue and Indian
    Revenue
87  gen log_movie_Indian_revenue = log(movie_Indian_Revenue)
88  gen log_movie_US_revenue = log(movie_US_Revenue)
89  gen log_movie_production_budget = log(movie_production_budget)
90  *** REPORT SUMMARY STATISTICS
91  summarize log_movie_Indian_revenue
92  summarize log_movie_US_revenue
93  summarize gender_ratio
94  summarize log_movie_production_budget
95  summarize i.startype
96  summarize i.movie_release_year
97  summarize i.movie_genre_Num
98  summarize i.movie_month_Num
99  summarize popular_revie
100 summarize critics_review
101 summarize movie_sequel
102 ***RUN INDIAN BOX OFFICE REGRESSION**
103 reg log_movie_Indian_revenue log_movie_production_budget
    gender_ratio i.startype i.movie_release_year i.movie_genre_Num i.
    movie_month_Num popular_review critics_review movie_sequel, robust
104 vif
105 estat imtest, white
106 *** RUN US BOX OFFICE REGRESSION*
107 reg log_movie_US_revenue log_movie_production_budget gender_ratio i
    .startype i.movie_release_year i.movie_genre_Num i.movie_month_Num
    popular_review critics_review movie_sequel, robust
108 vif
109 estat imtest, white
```