**Capstone Project: Appendix**

Sophia Balentine

Colorado State University Global

MIS581 - Capstone - Business Intelligence and Data Analytics

Dr. Osama Morad

August 3, 2025

**Github Link**

The following is the link to my Github Repository for this project which includes copies of all screenshots, code, presentations, and documents related to this project:

[**https://github.com/sophiabalentine27/MIS581\_Capstone\_StreamingProfits**](https://github.com/sophiabalentine27/MIS581_Capstone_StreamingProfits)

**All Project Code**

*Importing the subscriber number dataset:*

%web\_drop\_table(WORK.SERVICES);

FILENAME REFFILE '/home/u63708987/Capstone/Services.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.SERVICES;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.SERVICES; RUN;

*Creating a bar chart that shows the number of subscribers in the top six United States based streaming services platform*:

data attrmap;

length id $20 value $50 fillcolor $20;

id = "colorid1";

value = "Netflix"; fillcolor = "#E50914"; output;

value = "Amazon Prime Video"; fillcolor = "#00A8E1"; output;

value = "Disney+"; fillcolor = "#113CCF"; output;

value = "HBO Max"; fillcolor = "#5B2C91"; output;

value = "Hulu"; fillcolor = "#1CE783"; output;

value = "Apple TV+"; fillcolor = "#1D1D1F"; output;

run;

ods graphics / reset width=6in height=5in imagemap;

proc sgplot data=WORK.SERVICES(obs=6) dattrmap=attrmap;

where Service not in ("YouTube Premium", "Paramount+", "Peacock", "ESPN+", "Tencent Video", "iQIYI", "Youku", "ALTBalaji", "iflix");

vbar Service /

response=Subscribers

group=Service

groupdisplay=cluster

datalabel

attrid=colorid1;

xaxis label="Streaming Platform" discreteorder=data;

yaxis label="Number of Subscribers" grid;

title "Number of Subscribers per Streaming Platform";

run;

ods graphics / reset;

*Importing all of the streaming platform content tracker datasets:*

%web\_drop\_table(WORK.PRIME);

FILENAME REFFILE '/home/u63708987/Capstone/disney\_plus\_titles.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.PRIME;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.PRIME; RUN;

%web\_open\_table(WORK.PRIME);

%web\_drop\_table(WORK.DISNEY);

FILENAME REFFILE '/home/u63708987/Capstone/disney\_plus\_titles.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.DISNEY;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.DISNEY; RUN;

%web\_open\_table(WORK.DISNEY);

%web\_drop\_table(WORK.HBO);

FILENAME REFFILE '/home/u63708987/Capstone/HBO\_Content.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.HBO;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.HBO; RUN;

%web\_open\_table(WORK.HBO);

%web\_drop\_table(WORK.HULU);

FILENAME REFFILE '/home/u63708987/Capstone/hulu\_titles.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.HULU;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.HULU; RUN;

%web\_open\_table(WORK.HULU);

%web\_drop\_table(WORK.NETFLIX);

FILENAME REFFILE '/home/u63708987/Capstone/Netflix TV Shows and Movies.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.NETFLIX;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.NETFLIX; RUN;

%web\_open\_table(WORK.NETFLIX);

%web\_drop\_table(WORK.APPLE);

FILENAME REFFILE '/home/u63708987/Capstone/Apple titles.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.APPLE;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.APPLE; RUN;

%web\_open\_table(WORK.APPLE);

*Creating a table and bar chart to count the number of titles per platforms’ library:*

data lib\_vol;

length Platform $10;

if 0 then set netflix nobs=n\_netflix;

if 0 then set apple nobs=n\_apple;

if 0 then set prime nobs=n\_prime;

if 0 then set disney nobs=n\_disney;

if 0 then set hulu nobs=n\_hulu;

if 0 then set hbo nobs=n\_hbo;

Platform = "Netflix"; Count = n\_netflix; output;

Platform = "Apple TV+"; Count = n\_apple; output;

Platform = "Amazon Prime Video"; Count = n\_prime; output;

Platform = "Disney+"; Count = n\_disney; output;

Platform = "Hulu"; Count = n\_hulu; output;

Platform = "HBO Max"; Count = n\_hbo; output;

run;

data attrmap;

length id $20 value $50 fillcolor $20;

id = "colorid3";

value = "Netflix";fillcolor = "#E50914";output;

value = "Amazon Prime Video";fillcolor = "#00A8E1";output;

value = "Disney+";fillcolor = "#113CCF";output;

value = "HBO Max";fillcolor = "#5B2C91";output;

value = "Hulu";fillcolor = "#1CE783";output;

value = "Apple TV+";fillcolor = "#1D1D1F";output;

run;

proc sgplot data=lib\_vol dattrmap=attrmap;

vbar Platform /

response=Count

group=Platform

datalabel

attrid=colorid3;

title "Number of Observations per Streaming Platform";

run;

*Running a Correlation Analysis on amount of content offered and number of subscribers:*

data lib\_corr;

length Platform $25;

Platform = "Netflix";

LibraryVolume = 5440;

Subscribers = 221.8;

output;

Platform = "Apple TV+";

LibraryVolume = 170;

Subscribers = 20;

output;

Platform = "Amazon Prime Video";

LibraryVolume = 1450;

Subscribers = 175;

output;

Platform = "Disney+";

LibraryVolume = 1450;

Subscribers = 129.8;

output;

Platform = "Hulu";

LibraryVolume = 3090;

Subscribers = 45.3;

output;

Platform = "HBO Max";

LibraryVolume = 1386;

Subscribers = 73.8;

output;

run;

ods noproctitle;

ods graphics / imagemap=on;

proc corr data=lib\_corr pearson plots=matrix;

var LibraryVolume Subscribers;

run;

*Importing the dataset with Rotten Tomato review scores from critics and viewers on all original television content produced by major streaming services:*

%web\_drop\_table(WORK.RTScores);

FILENAME REFFILE '/home/u63708987/Capstone/RT Scores.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.RTScores;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.RTScores; RUN;

%web\_open\_table(WORK.RTScores);

*Cleaning the Rotten Tomatoes dataset to include only the services we want to look at and make the ratings numeric:*

data cleanedRTS;

set work.rtscores;

if Network not in ("VUDU", "Peacock", "Paramount");

CriticRating=input(compress('Critic Score'n, '%'), 8.) / 100;

ViewerRating=input(compress('Critic Score'n, '%'), 8.) / 100;

ViewPct= ViewerRating\*100;

Run;

*Creating a histogram of all the viewer ratings for each streaming platform:*

data attrmap;

length id $20 value $50 fillcolor $20;

id = "colorid2";

value = "Netflix"; fillcolor = "#E50914"; output;

value = "Prime Vid"; fillcolor = "#00A8E1"; output;

value = "Disney+"; fillcolor = "#113CCF"; output;

value = "HBO MAX"; fillcolor = "#5B2C91"; output;

value = "Hulu"; fillcolor = "#1CE783"; output;

value = "Apple TV+"; fillcolor = "#1D1D1F"; output;

run;

ods graphics / reset width=10in height=5in imagemap;

proc sgpanel data=work.cleanedrts dattrmap=attrmap;

panelby Network / columns=6 spacing=0 novarname;

histogram ViewPct /

binwidth=10

scale=percent

datalabel

group=Network

attrid=colorid2;

colaxis display=all

label="Viewer Rating (in %)"

values=(0 to 100 by 10);

rowaxis display=all

label="Frequency";

title "Histogram of Viewer Ratings by Streaming Service";

run;

*Creating a line chart to view the ratings from the histogram side by side:*

ods graphics / reset;

proc format;

value binfmt

0 - <10 = '0-9'

10 - <20 = '10-19'

20 - <30 = '20-29'

30 - <40 = '30-39'

40 - <50 = '40-49'

50 - <60 = '50-59'

60 - <70 = '60-69'

70 - <80 = '70-79'

80 - <90 = '80-89'

90 - 100 = '90-100';

run;

data binned;

set work.cleanedrts;

format bin binfmt.;

bin = floor(ViewPct / 10) \* 10;

if bin > 100 then bin = 100;

run;

proc sql;

create table binned\_summary as

select

Network,

catx('-', put(bin, 2.), put(bin+9, 2.)) as RatingBin,

bin as bin\_numeric,

count(\*) as Count

from binned

group by Network, bin

order by Network, bin;

quit;

proc sql;

create table line\_data as

select

a.Network,

a.RatingBin,

a.bin\_numeric,

a.Count,

(a.Count / b.Total)\*100 as Percent format=5.1

from binned\_summary a

inner join (

select Network, sum(Count) as Total

from binned\_summary

group by Network

) b

on a.Network = b.Network;

quit;

data attrmap;

length id $20 value $50 linecolor $20;

id = "colorid3";

value = "Netflix";linecolor = "#E50914";output;

value = "Prime Vid";linecolor = "#00A8E1";output;

value = "Disney+";linecolor = "#113CCF";output;

value = "HBO MAX";linecolor = "#5B2C91";output;

value = "Hulu";linecolor = "#1CE783";output;

value = "Apple TV+";linecolor = "#1D1D1F";output;

run;

ods graphics / reset width=10in height=6in;

proc sgplot data=line\_data dattrmap=attrmap;

series x=bin\_numeric y=Percent /

group=Network

lineattrs=(thickness=2)

attrid=colorid3;

xaxis label="Viewer Rating (%)" values=(0 to 100 by 10);

yaxis label="Percentage of Ratings";

title "Viewer Rating Distribution by Streaming Service";

run;

ods graphics / reset;

*Running an ANOVA analysis on viewer ratings and number of subscribers*

Title;

ods noproctitle;

ods graphics / imagemap=on;

proc glm data=WORK.IMPORT;

class Subscribers;

model 'Number Score'n=Subscribers;

means Subscribers / hovtest=levene welch plots=none;

lsmeans Subscribers / adjust=tukey pdiff alpha=.05;

run;

quit;

*Importing the User Churn in telecommunications simulated dataset:*

%web\_drop\_table(WORK.CHURN);

FILENAME REFFILE '/home/u63708987/Capstone/User churn.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.CHURN;

GETNAMES=YES;

RUN;

PROC CONTENTS DATA=WORK.CHURN; RUN;

%web\_open\_table(WORK.CHURN);

*Cleaning the User Churn dataset so that it is usable:*

data churn\_clean;

set work.CHURN;

if Churn = "Yes" and (StreamingTV = "Yes" or StreamingMovies = "Yes");

run;

*Creating a bar chart of high churn users’ contract types:*

proc sgplot data=churn\_clean;

vbar Contract / datalabel;

title "Contract Types of Churn Customers who Utilize Streaming";

run;

*Running a Correlation Analysis on tenure of customer versus monthly charges:*

ods noproctitle;

ods graphics / imagemap=on;

proc corr data=WORK.CHURN\_CLEAN pearson nosimple noprob plots=matrix;

var MonthlyCharges;

with tenure;

run;

*Running an ANOVA analysis on tenure and contract type:*

ods noproctitle;

ods graphics / imagemap=on;

proc corr data=WORK.CHURN\_CLEAN pearson nosimple noprob plots=matrix;

var MonthlyCharges;

with tenure;

run;