Institute of Systems Science

National University of Singapore

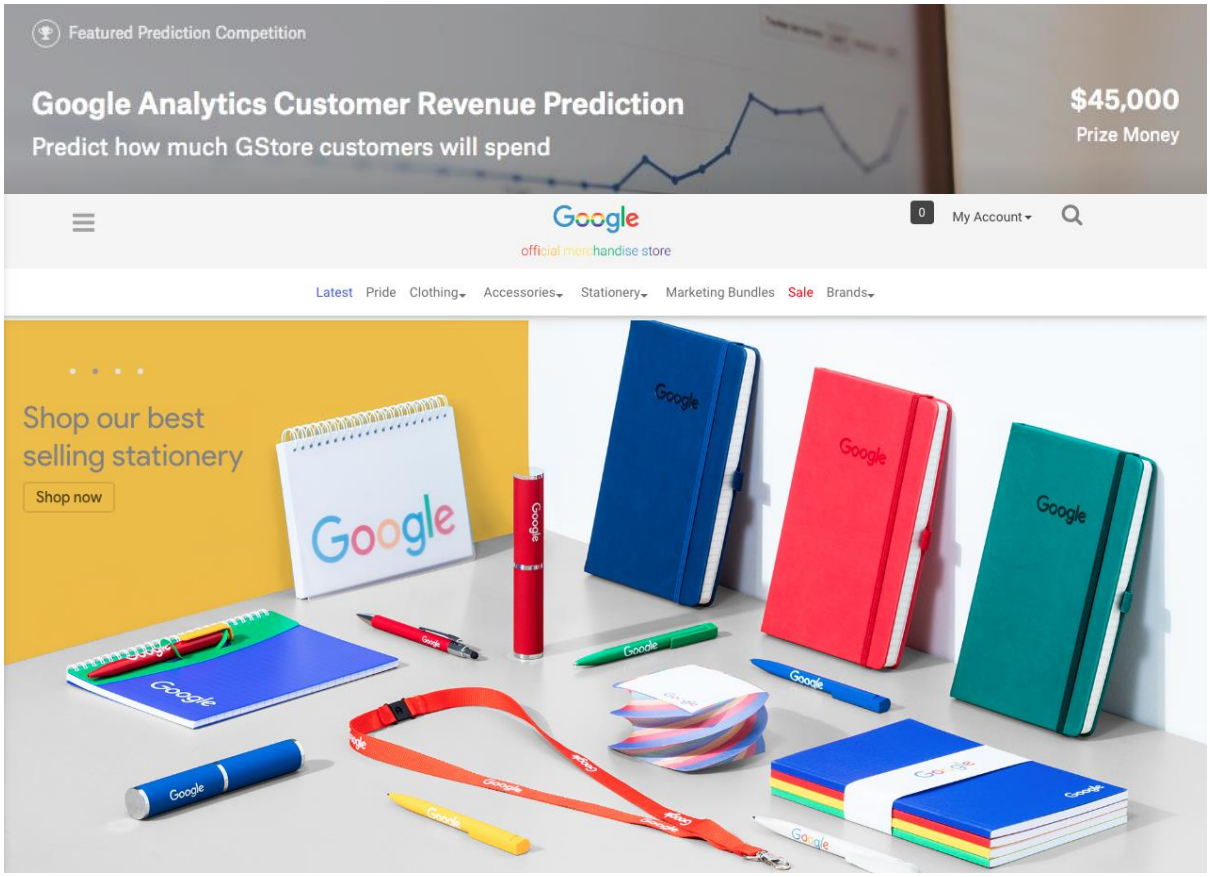
GRADUATE CERTIFICATE

BUSINESS ANALYTICS PRACTICE

Supplementary Workshop Guide

Subject: *NICF- Statistics Bootcamp (SF)*

Workshop: Exploratory Data Analytics (EDA)
&
Predictive Modeling



Expected learning outcomes

Knowledge:

- Analyse a case of customer revenue for Google Merchandise Store, using Google Analytics data.
- Understand the various features available from Google Analytics data.

Abilities:

- Able to conduct exploratory data analysis (EDA) using ***R***
- Able to use predictive models to forecast future customer revenue using ***R***
- Able to identify important features which can indicate customer purchases (data insights)
- Able to build interactive data visualization dashboard to present results & insights using ***Tableau***

Case Study: Google Analytics Customer Revenue Prediciton

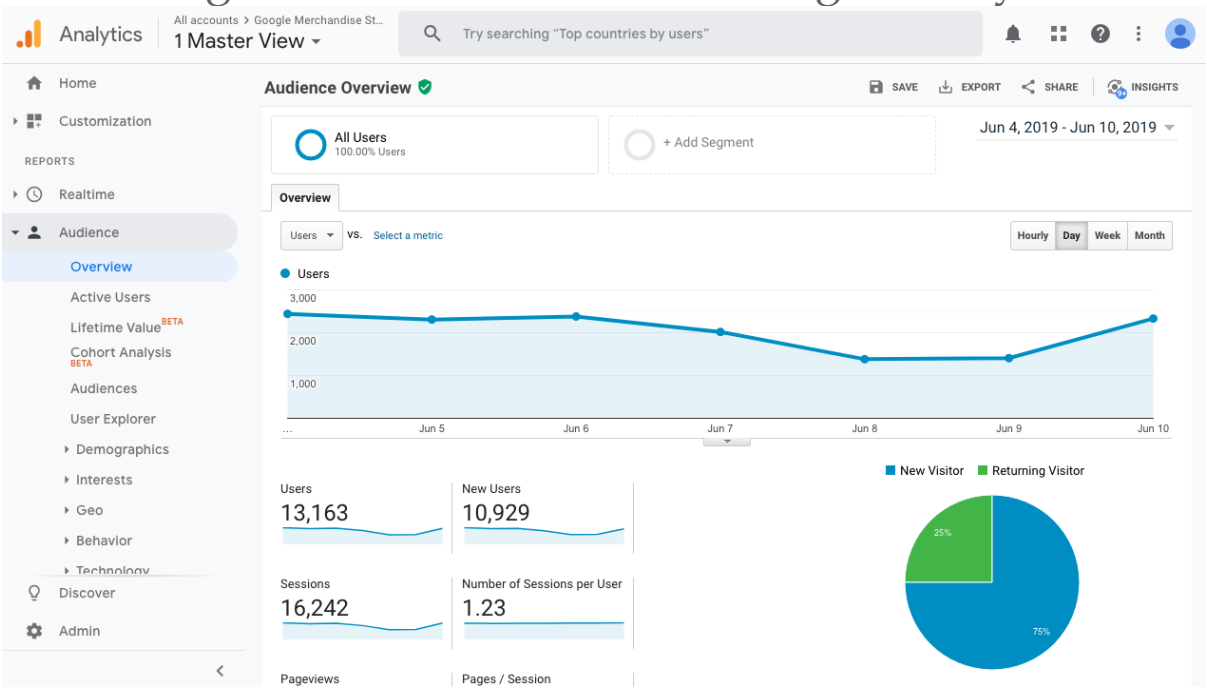
https://www.kaggle.com/c/ga-customer-revenue-prediction

The 80/20 rule has proven true for many businesses–only a small percentage of customers produce most of the revenue. As such, marketing teams are challenged to make appropriate investments in promotional strategies.

RStudio, the developer of free and open tools for R and enterprise-ready products for teams to scale and share work, has partnered with Google Cloud and Kaggle to demonstrate the business impact that thorough data analysis can have.
In this competition, you're challenged to analyze a Google Merchandise Store (also known as GStore, where Google swag is sold) customer dataset to predict revenue per customer. Hopefully, the outcome will be more actionable operational changes and a better use of marketing budgets for those companies who choose to use data analysis on top of GA data.

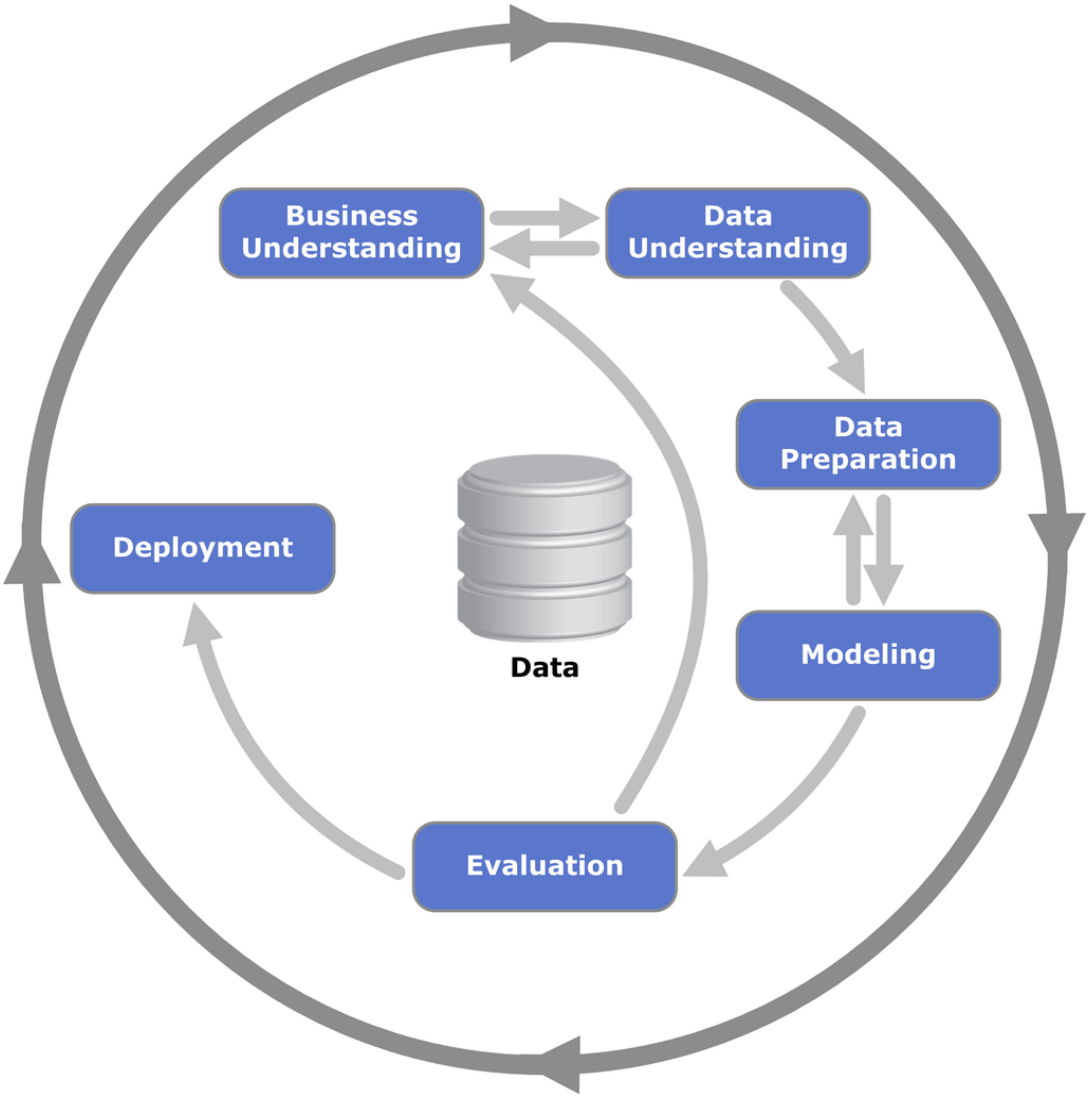
https://www.googlemerchandisestore.com/

Google Merchandise Store – Google Analytics



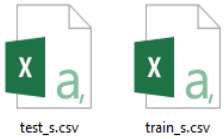
https://analytics.google.com/analytics/web/demoAccount

Data analytics life cycle



https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Data Understanding / Data Fields (Csv Files)



channelGrouping	customDimensions	date	device	fullVisitorId	geoNetwork	hits	socialEngagementType	totals	trafficSource	visitId	visitNumber	visitStartTime
Organic Search	[{'index': '4', 'value': 'North America'}]	20171016	{"browser": "Safari", "b	5.20065E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "7"	{"campaign": "(not set)	1508169977	2	1508169977
Organic Search	[]	20171016	{"browser": "Chrome",	9.06692E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1508152302	1	1508152302
Organic Search	[{'index': '4', 'value': 'APAC'}]	20171016	{"browser": "Chrome",	8.78191E+18	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "3"	{"campaign": "(not set)	1508154608	1	1508154608
Organic Search	[{'index': '4', 'value': 'EMEA'}]	20171016	{"browser": "Safari", "b	2.68366E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "2"	{"campaign": "(not set)	1508190767	1	1508190767
Display	[]	20171016	{"browser": "Chrome",	9.9269E+18	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "1000557	1508179547	2	1508179547
Direct	[{'index': '4', 'value': 'North America'}]	20171016	{"browser": "Chrome",	4.44331E+17	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "2"	{"campaign": "(not set)	1508190887	2	1508190887
Organic Search	[{'index': '4', 'value': 'North America'}]	20171016	{"browser": "Safari", "b	4.55907E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1508191209	1	1508191209
Organic Search	[{'index': '4', 'value': 'North America'}]	20171016	{"browser": "Chrome",	1.59165E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "2"	{"campaign": "(not set)	1508161244	1	1508161244
Organic Search	[]	20160902	{"browser": "Internet E	4.44545E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1472805784	1	1472805784
Affiliates	[{'index': '4', 'value': 'EMEA'}]	20160902	{"browser": "Chrome",	9.17E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "Data Sha	1472808470	3	1472808470
Organic Search	[]	20160902	{"browser": "Edge", "br	4.7004E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1472885051	1	1472885051
Organic Search	[{'index': '4', 'value': 'APAC'}]	20160902	{"browser": "Chrome",	5.11295E+17	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "3"	{"campaign": "(not set)	1472817112	5	1472817112
Social	[]	20160902	{"browser": "Chrome",	7.2092E+18	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"referralPath": "/yt/ab	1472808082	1	1472808082
Social	[{'index': '4', 'value': 'EMEA'}]	20160902	{"browser": "Safari", "b	4.0837E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"referralPath": "/yt/ab	1472818309	1	1472818309
Social	[{'index': '4', 'value': 'EMEA'}]	20160902	{"browser": "Firefox", "	4.58911E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "2"	{"referralPath": "/yt/ab	1472855774	1	1472855774
Paid Search	[{'index': '4', 'value': 'North America'}]	20171130	{"browser": "Chrome",	9.6631E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "5"	{"campaign": "(not set)	1512078847	11	1512078847
Organic Search	[{'index': '4', 'value': 'North America'}]	20171130	{"browser": "Safari", "b	7.38885E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "9"	{"campaign": "(not set)	1512072709	2	1512072709
Organic Search	[{'index': '4', 'value': 'EMEA'}]	20171130	{"browser": "Safari", "b	9.61903E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1512068072	3	1512068072
Direct	[]	20171130	{"browser": "Internet E	1.15056E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1512094818	1	1512094818
Direct	[{'index': '4', 'value': 'North America'}]	20171130	{"browser": "Chrome",	1.12545E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1512079310	4	1512079310
Referral	[]	20171130	{"browser": "Chrome",	2.50595E+18	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "3"	{"referralPath": "/analy	1512103166	1	1512103166
Affiliates	[{'index': '4', 'value': 'EMEA'}]	20171130	{"browser": "Opera", "b	8.93746E+17	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "4"	{"campaign": "Data Sha	1512048656	1	1512048656
Direct	[]	20171130	{"browser": "Chrome",	7.18328E+18	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1512054683	1	1512054683
Organic Search	[{'index': '4', 'value': 'APAC'}]	20171130	{"browser": "Safari", "b	1.19683E+16	{"continent": "Asia", "s	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1512050597	1	1512050597
Organic Search	[{'index': '4', 'value': 'EMEA'}]	20171130	{"browser": "Chrome",	4.31198E+17	{"continent": "Europe",	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "3"	{"campaign": "(not set)	1512075439	1	1512075439
Organic Search	[{'index': '4', 'value': 'North America'}]	20170126	{"browser": "Chrome",	3.52554E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "1"	{"campaign": "(not set)	1485452557	10	1485452557
Direct	[{'index': '4', 'value': 'North America'}]	20170126	{"browser": "Chrome",	9.25318E+16	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "4"	{"campaign": "(not set)	1485486313	19	1485486313
Organic Search	[{'index': '4', 'value': 'North America'}]	20170126	{"browser": "Safari", "b	6.36149E+18	{"continent": "America	[{'hitNumber': '1', 'time	Not Socially Engaged	{"visits": "1", "hits": "11	{"campaign": "(not set)	1485489966	1	1485489966

- ***channelGrouping*** - The channel via which the user came to the Store.
- ***customDimensions*** - This section contains any user-level or session-level custom dimensions that are set for a session. This is a repeated field and has an entry for each dimension that is set.
- ***date*** - The date on which the user visited the Store.
- ***device*** - The specifications for the device used to access the Store.
- ***fullVisitorId***- A unique identifier for each user of the Google Merchandise Store.
- ***geoNetwork*** - This section contains information about the geography of the user.
- ***hits*** - This row and nested fields are populated for any and all types of user behaviours (hits). Provides a record of all page visits.
- ***socialEngagementType*** - Engagement type, either "Socially Engaged" or "Not Socially Engaged".
- ***totals*** - This set of columns mostly includes high-level aggregate data, including target variable: transactionRevenue, e.g. one sample field value of 'totals':

  {
  "visits" : "1",
  "hits" : "16",
  "pageviews" : "15",
  "timeOnSite" : "225",
  "transactions" : "1",
  "transactionRevenue" : "21990000",
  "newVisits" : "1",
  "totalTransactionRevenue" : "29990000",
  "sessionQualityDim" : "6"
  }

- ***trafficSource*** - This section contains information about the Traffic Source from which the session originated.
- ***visitId*** - An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
- ***visitNumber*** - The session number for this user. If this is the first session, then this is set to 1.
- ***visitStartTime*** - The timestamp (expressed as POSIX time).

What to predict ?

We are predicting the **natural log** of the sum of all transactions revenue **per customer**. For every customer in the test data set, the target is:

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

Evaluation: Root Mean Squared Error (RMSE)

Submissions are scored on the root mean squared error. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y hat is the natural log of the predicted summed transaction revenue for a customer and y is the natural log of the actual summed transaction revenue value plus one.

Analytics using software

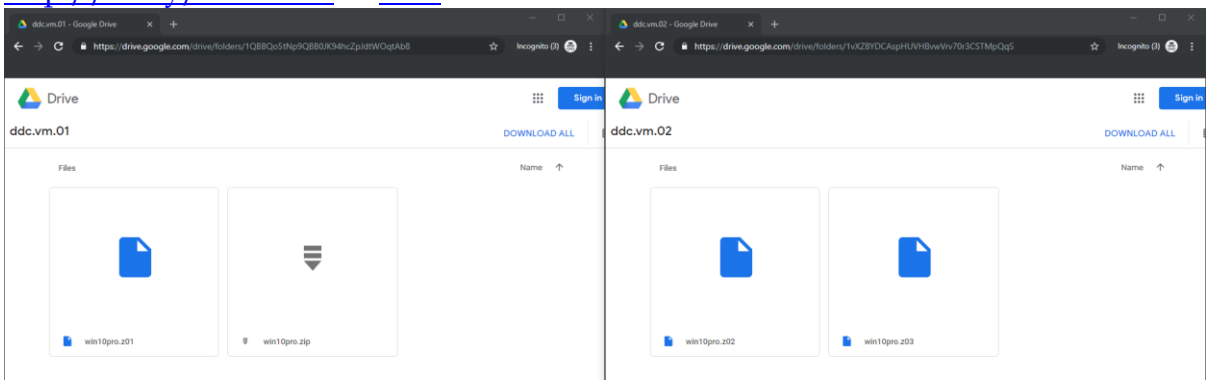
Download & install Software: *VirtualBox 5.2.20*
https://download.virtualbox.org/virtualbox/5.2.20/VirtualBox-5.2.20-125813-Win.exe

Also Download & install: *VirtualBox 5.2.20 Extension Pack*
https://download.virtualbox.org/virtualbox/5.2.20/Oracle_VM_VirtualBox_Extension_Pack-5.2.20.vbox-extpack

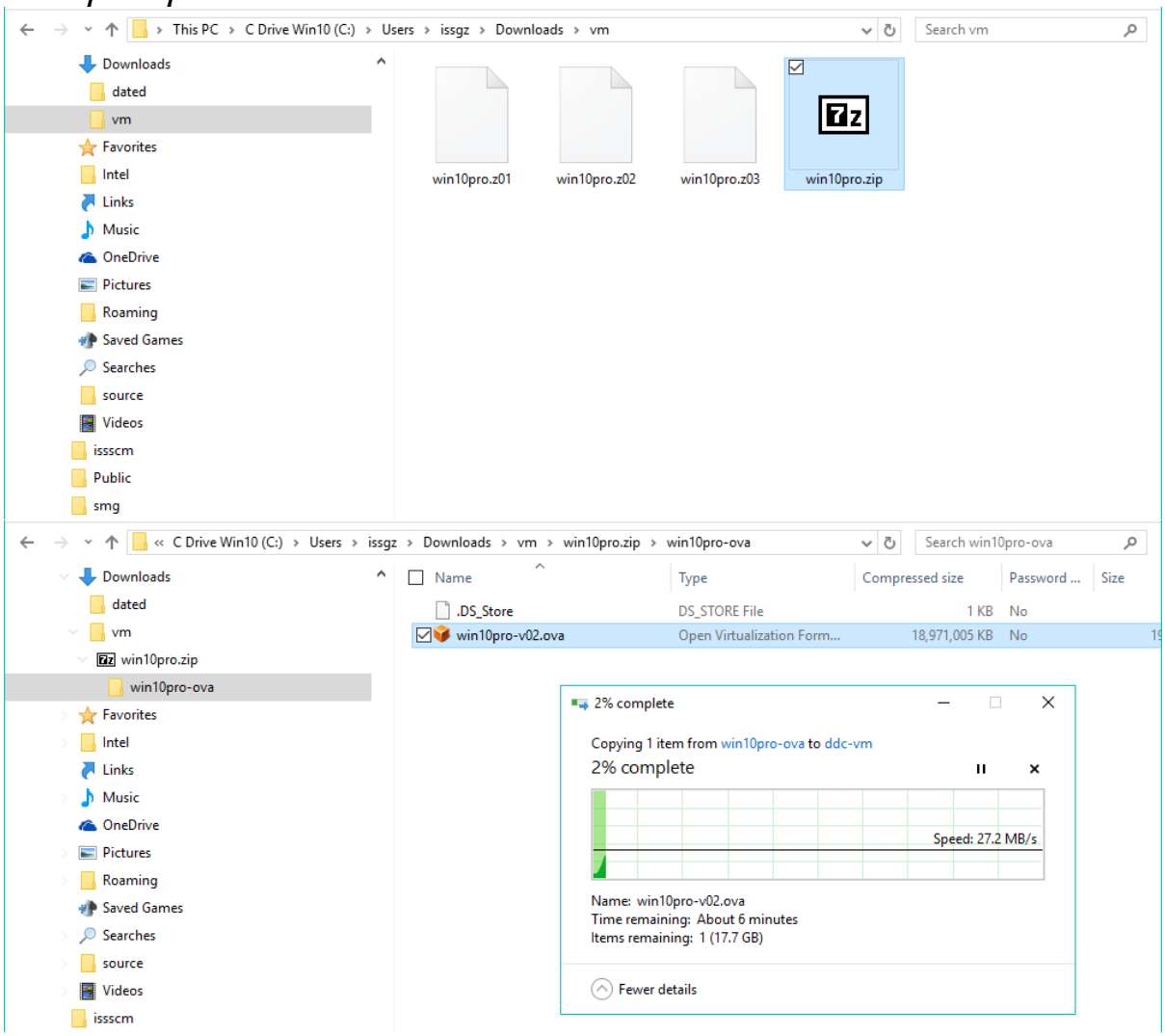
Download virtual machine workstation: *win10pro*

Download four VM files (split zip file containing win10pro.ova image) from below two links:

http://bit.ly/ddcvm01 or here
http://bit.ly/ddcvm02 or here

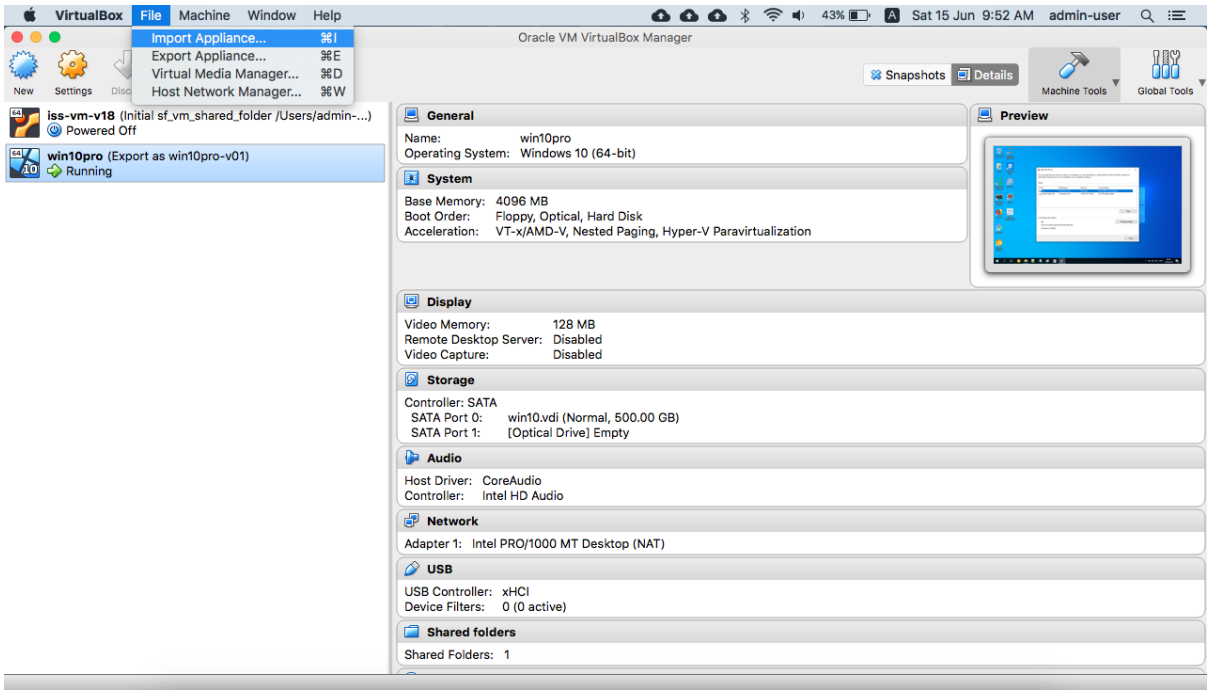


Save all files into **same** file folder; select to unzip the master zip file named: ***win10pro.zip***.

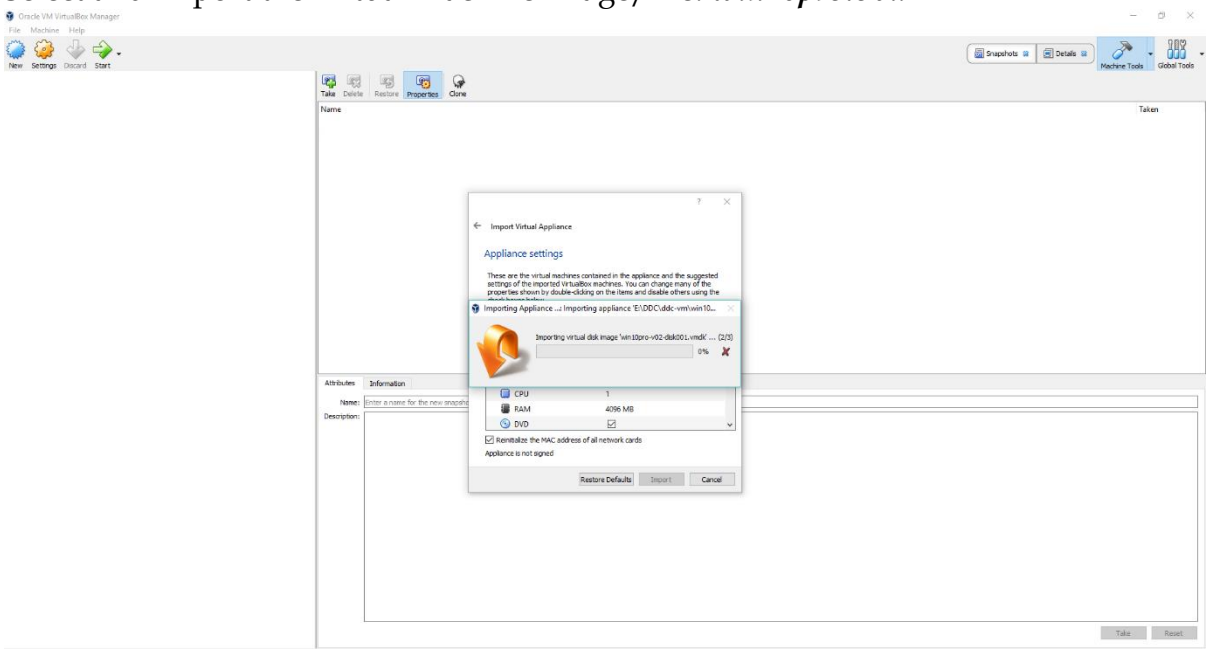


Import & run virtual machine workstation: *win10pro*

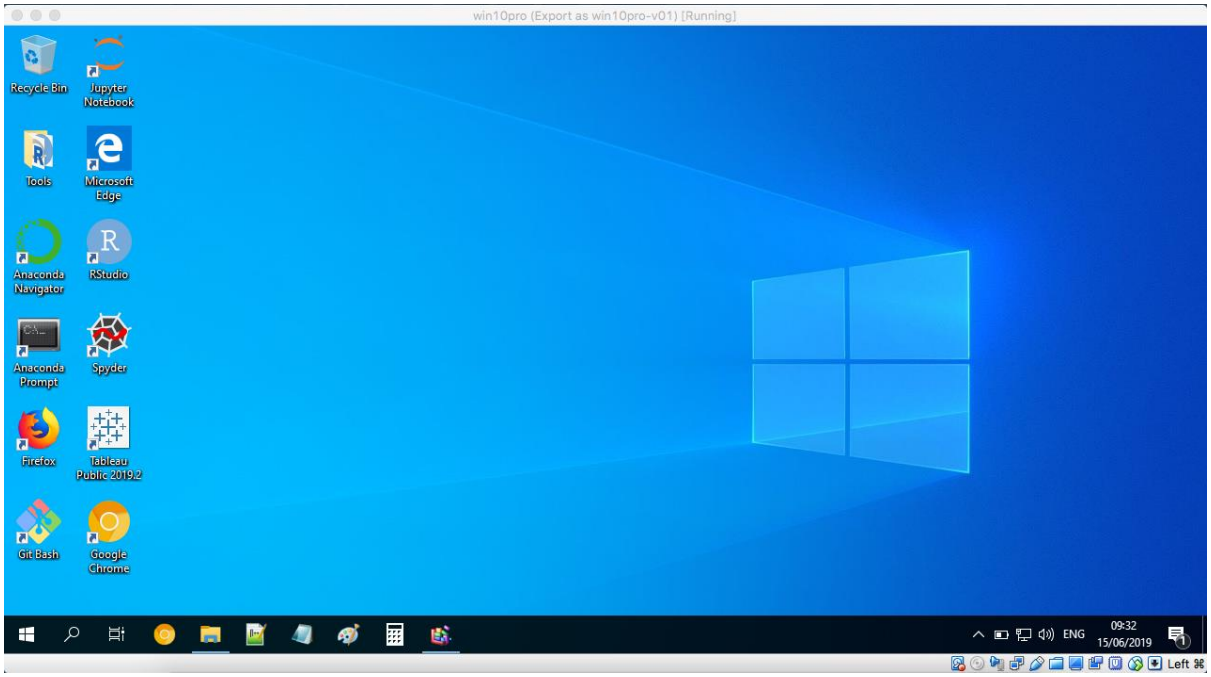
Start VirtualBox, click ***File → Import Appliance***.



Select and import the virtual machine image/file: ***win10pro.ova***



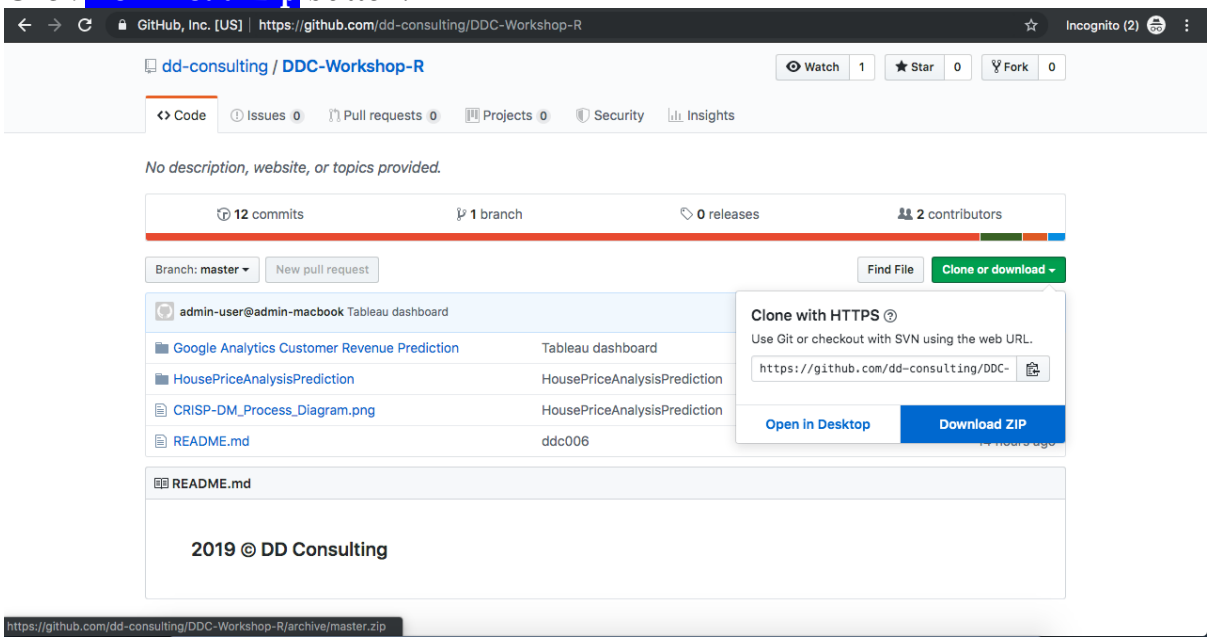
Click show  to start virtual machine.



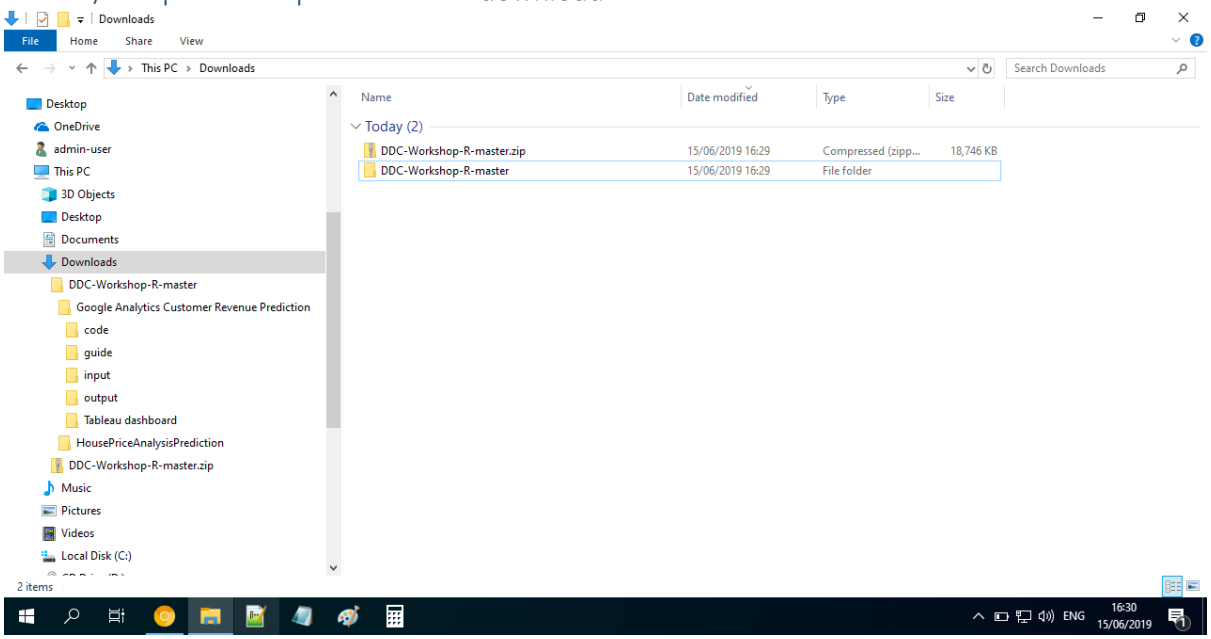
Download workshop materials

Inside win10pro virtual machine, open web browser: https://github.com/dd-consulting/DDC-Workshop-R

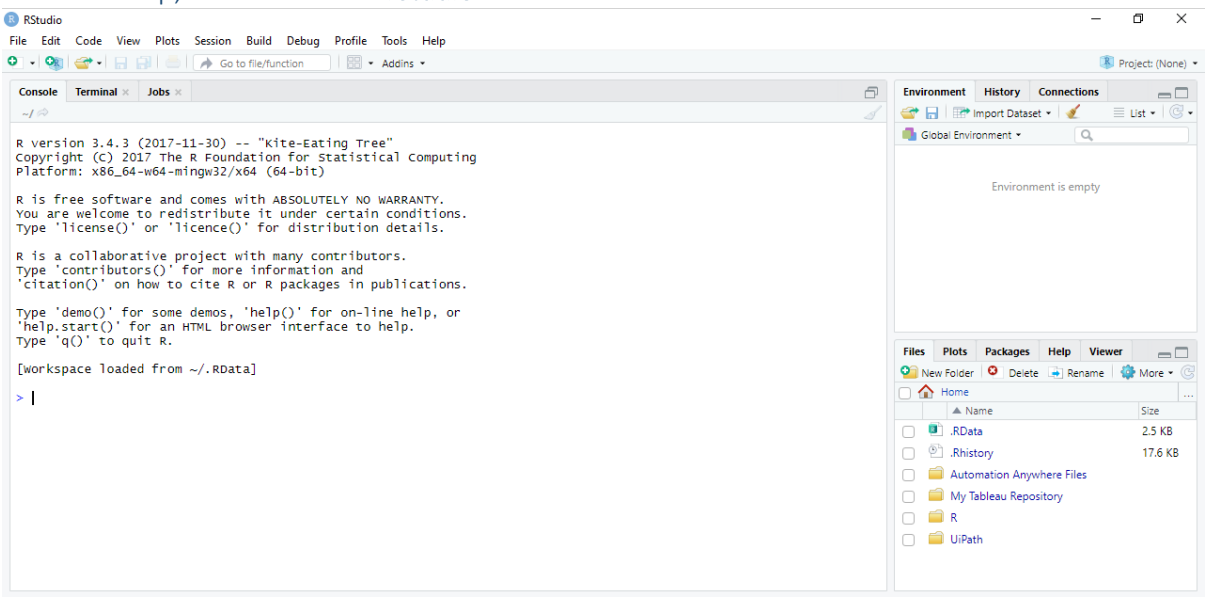
Click **Download Zip** button.



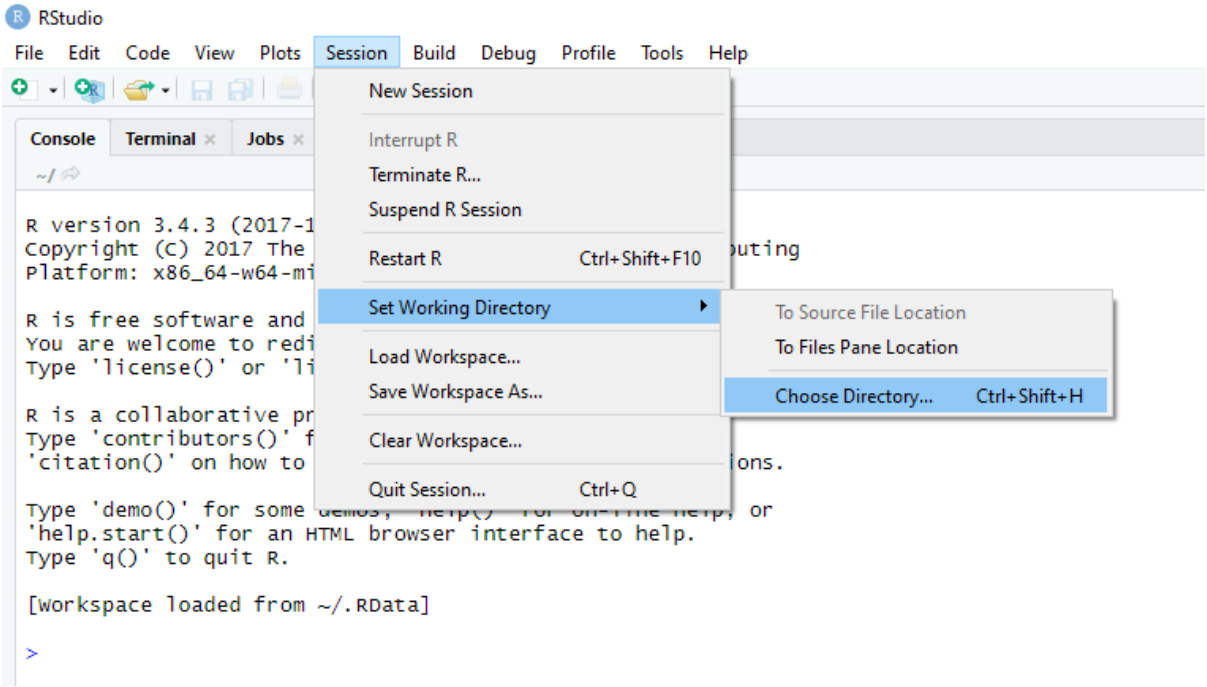
Extract/Unzip workshop materials to ***download*** folder



from desktop, Strat software: *rstudio*

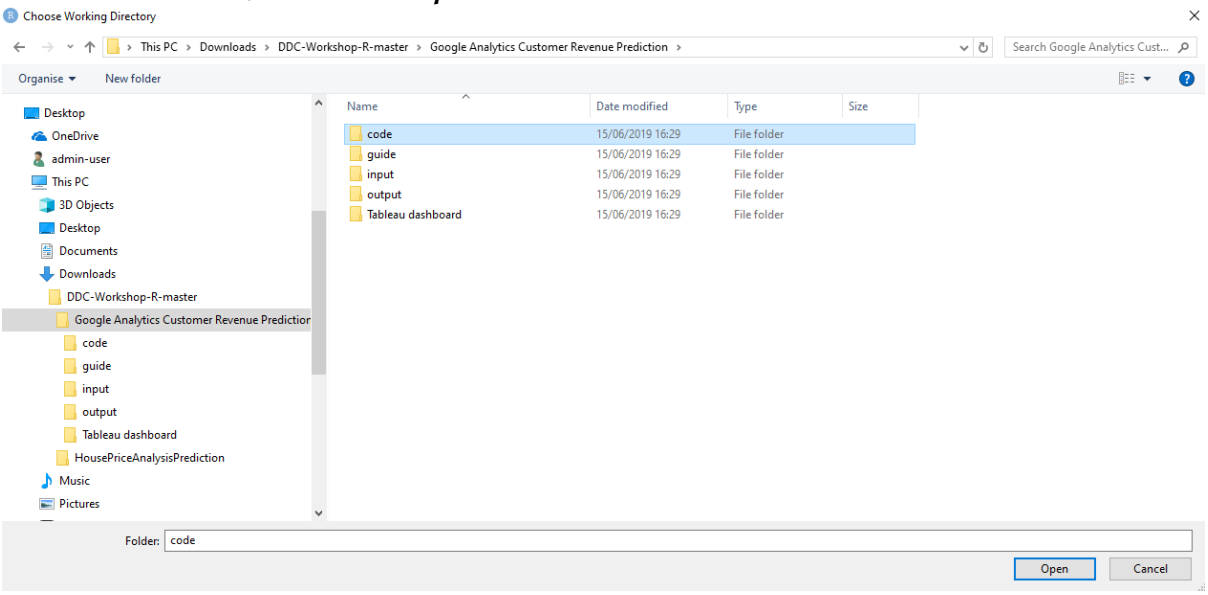


set working directory



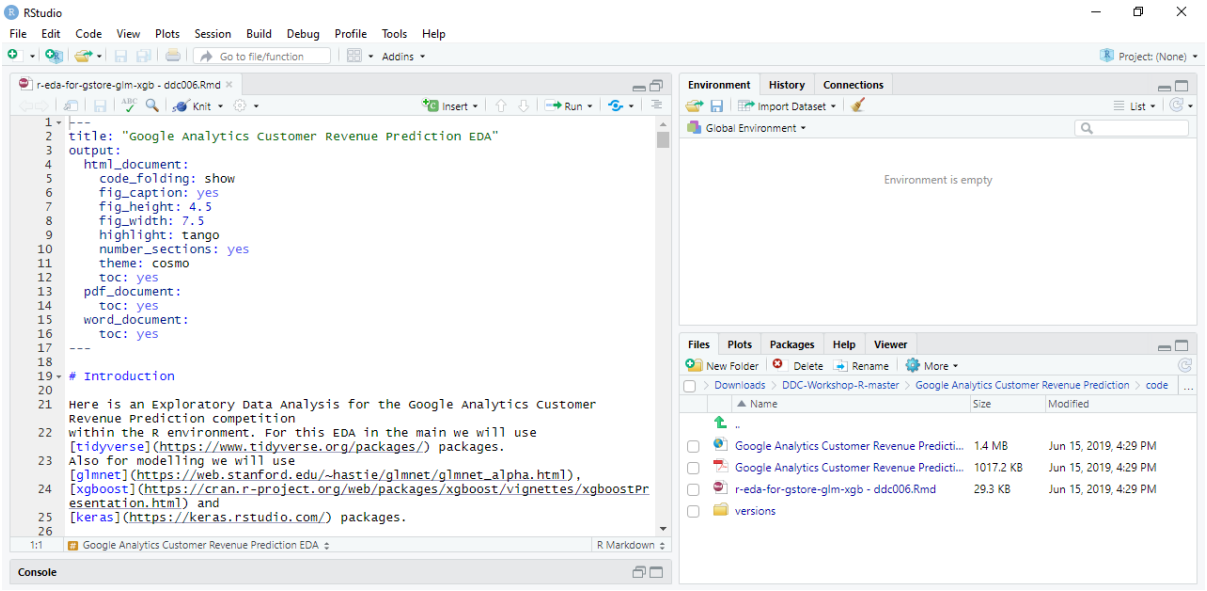
C:\Users\admin-user\Downloads\DDC-Workshop-R-master\Google Analytics Customer Revenue Prediction\code

Select folder ***code***; then click ***Open***

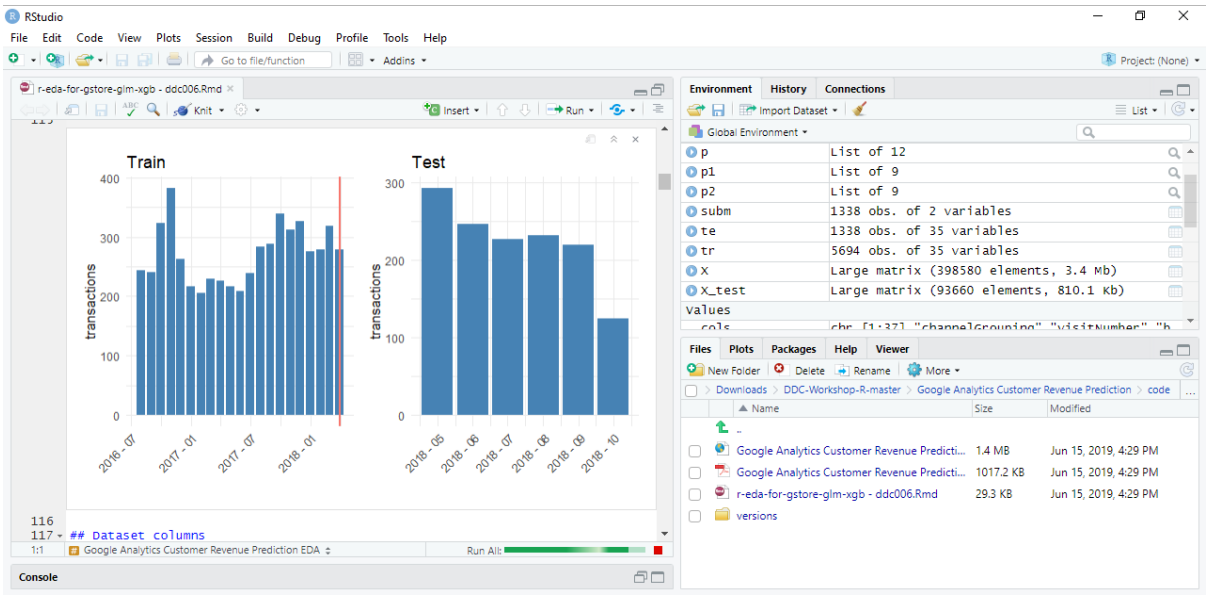


Open R Markdown/scrpt file

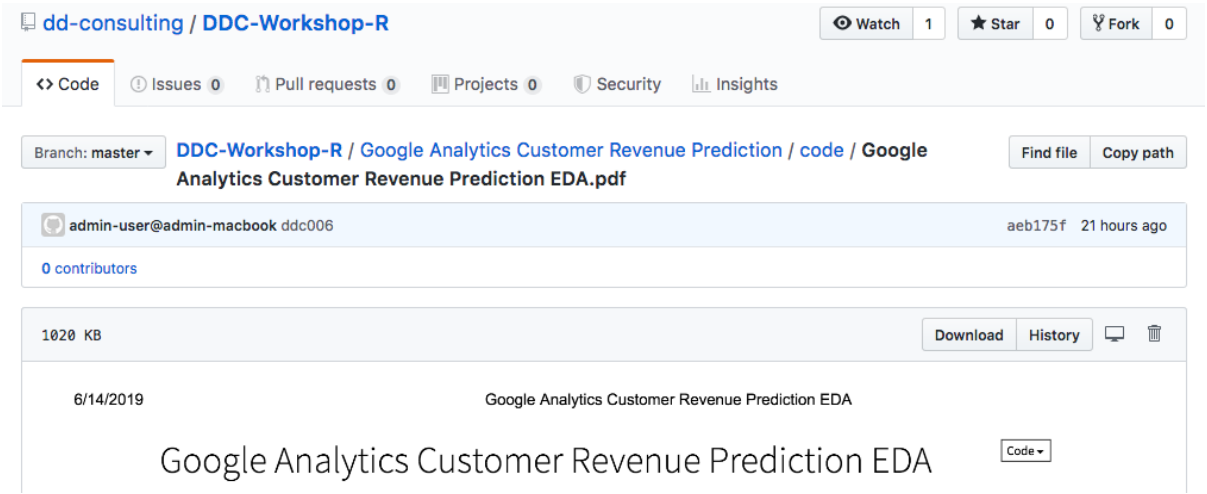
E.g. Google Analytics Customer Revenue Prediction/code/r-eda-for-gstore-glm-xgb – ddc006.Rmd



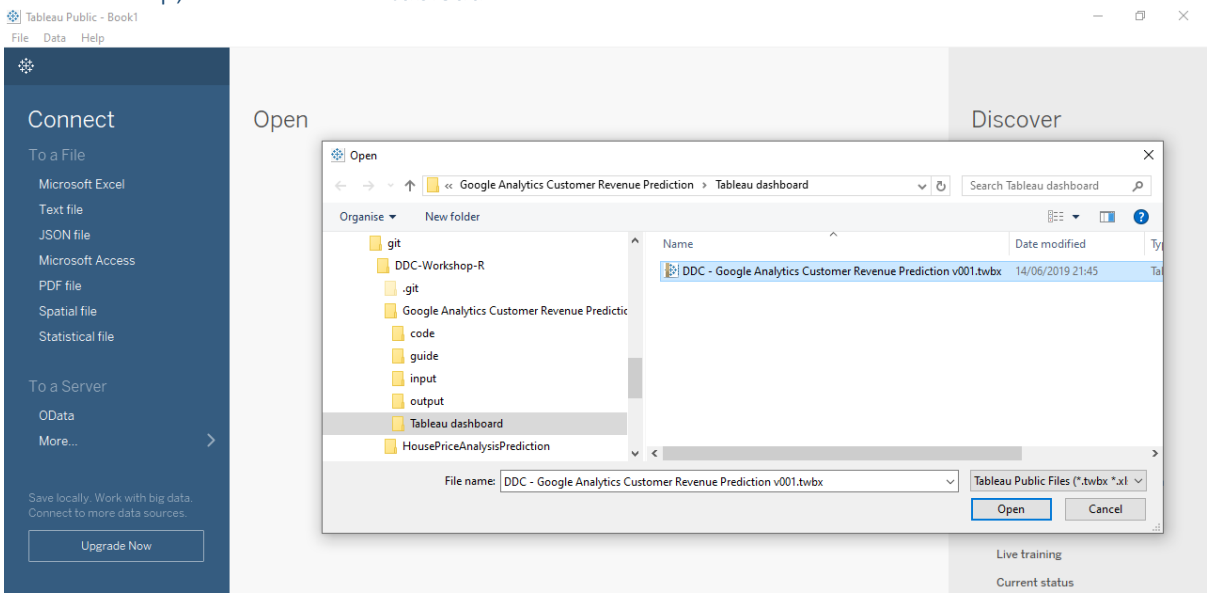
Run script for EXPLORATORY Data Analysis (EDA) & Predictive Modelling



PDF file exported from R Markdown: https://github.com/dd-consulting/DDC-Workshop-R/tree/master/Google%20Analytics%20Customer%20Revenue%20Prediction/code

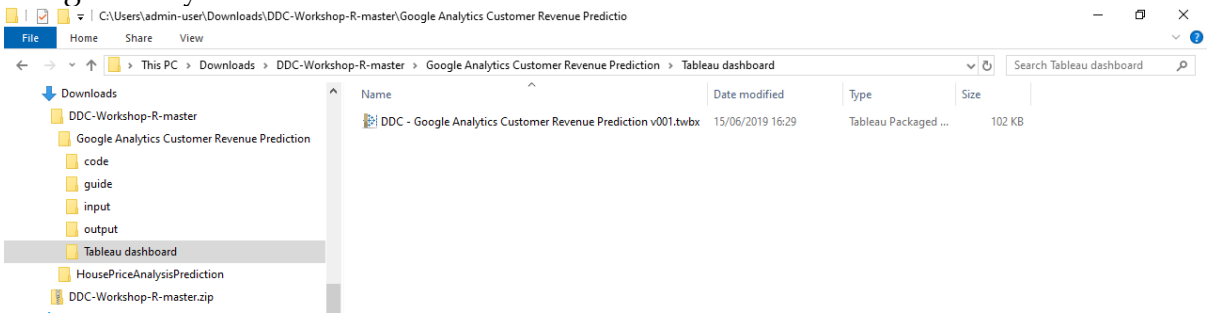


from desktop, Strat software: *tableau*



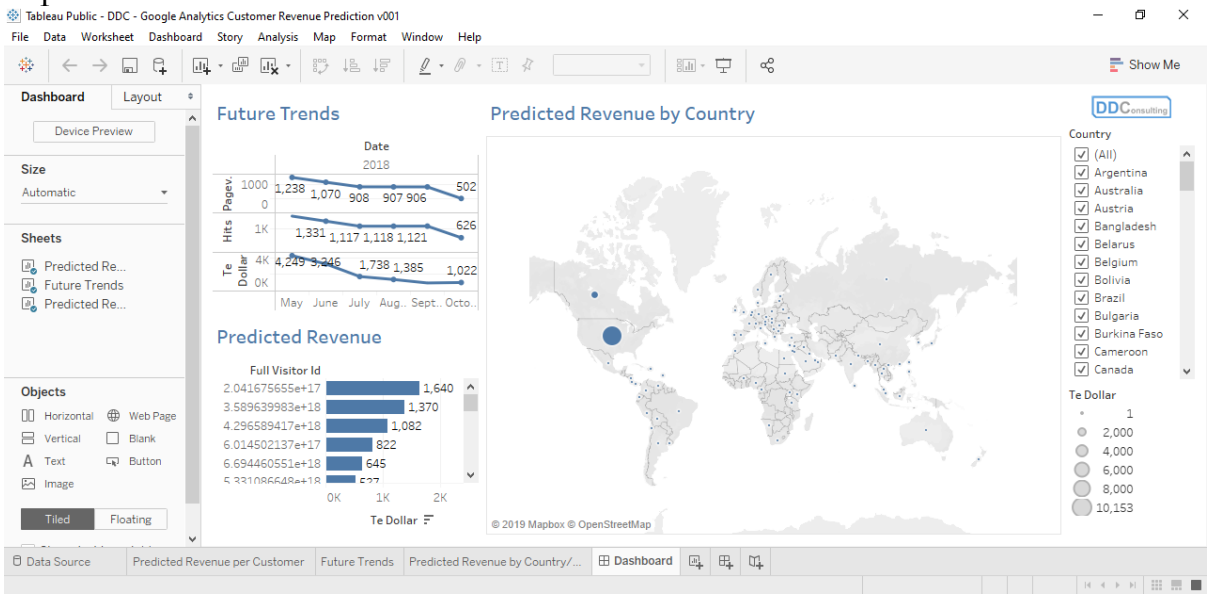
Open Tableau file: Interactive Data visualization

E.g. Google Analytics Customer Revenue Prediction/Tableau dashboard/DDC - Google Analytics Customer Revenue Prediction.twbx



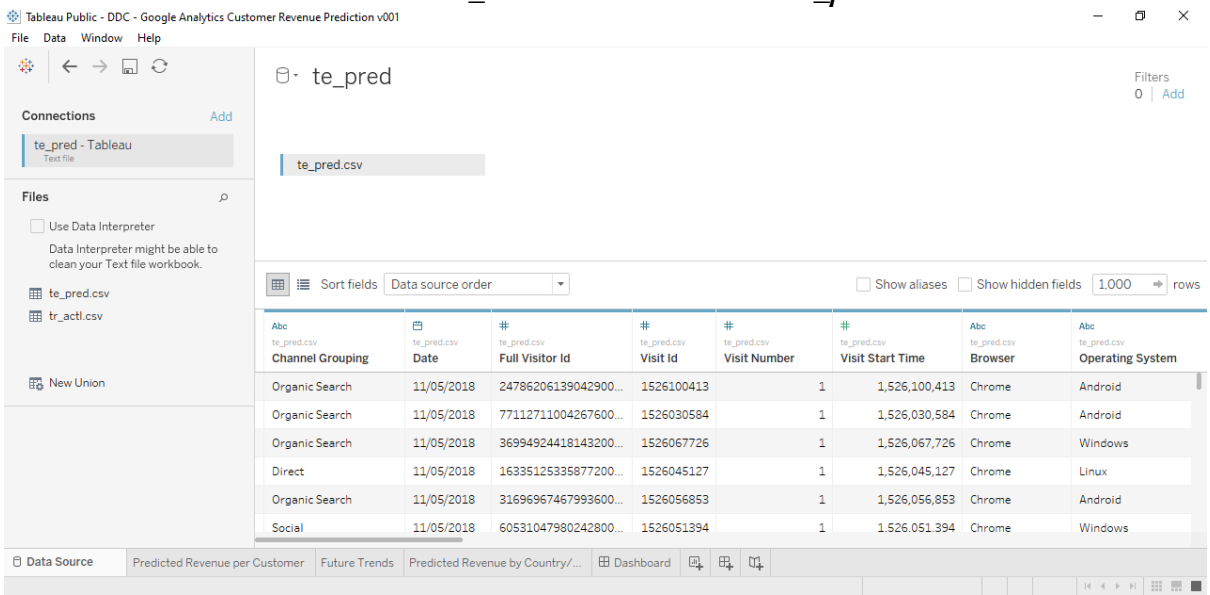
explore, edit, and Create Tableau visualizations

Explore to below three ***Worksheets*** and one ***Dashboard***:



Challenge you: Enhance the visualizations where applicable; or replicate the ***Worksheets*** or ***Dashboard*** of your choice.

Make use of new ***Data Source***: ***tr_actl – Tablaeu.csv*** or ***te_pred – Tablaeu.csv***

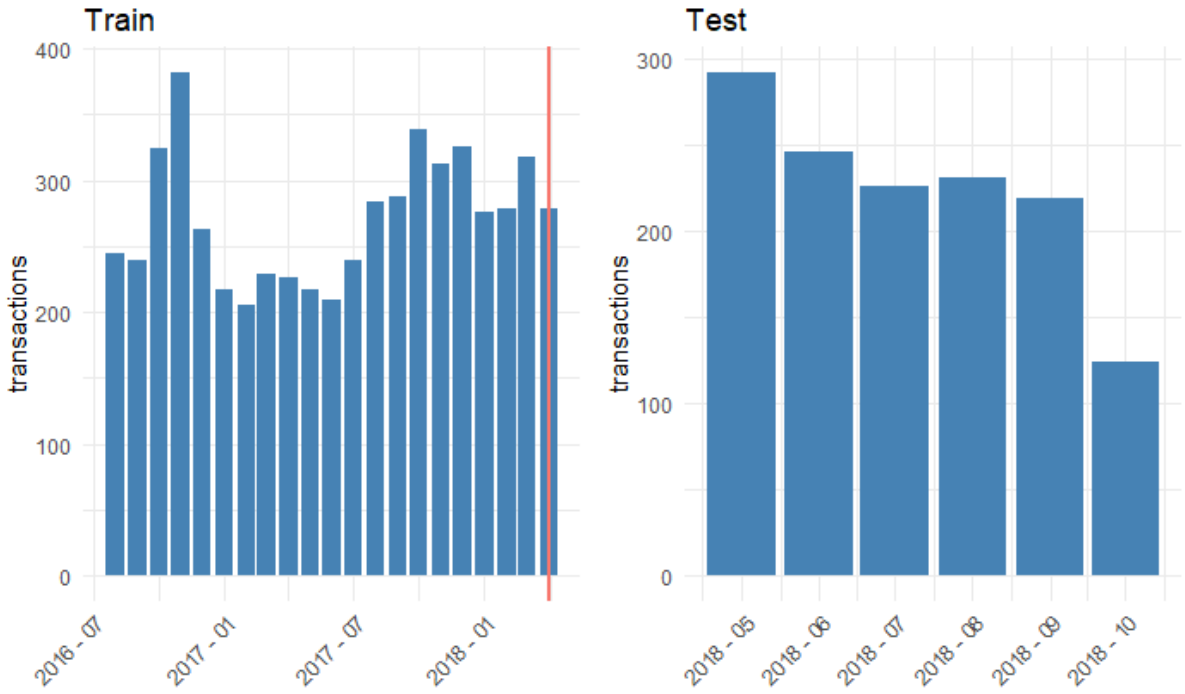


Online data visualization reference:
https://public.tableau.com/profile/dd.consulting#!/

Workshop Exercises

Question 1 : How many **months** of data are there in training data?
Your answer :

Question 2 : How many **months** of data are there in testing data?
Your answer :

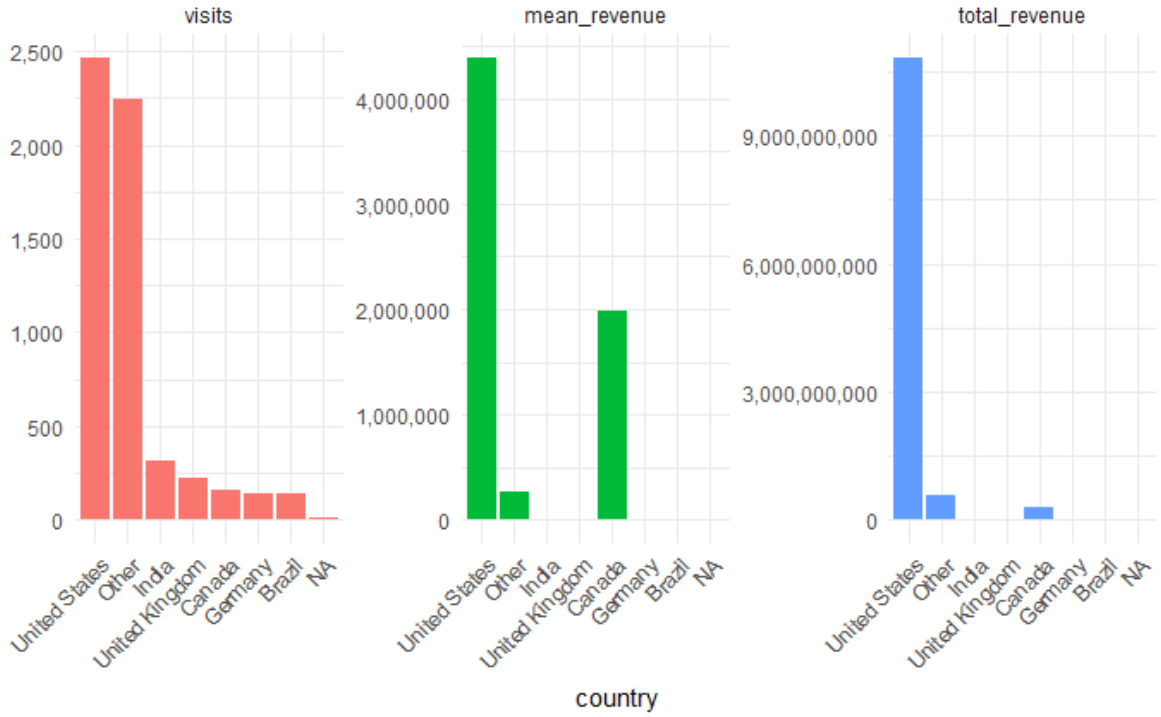


Question 3 : What's the **variable name** of the target customer revenue we want to predict?
Your answer :

```
 [1] "channelGrouping"               "date"
 [3] "fullVisitorId"                 "visitId"
 [5] "visitNumber"                   "visitStartTime"
 [7] "browser"                       "operatingSystem"
 [9] "isMobile"                      "deviceCategory"
[11] "continent"                     "subContinent"
[13] "country"                       "region"
[15] "metro"                         "city"
[17] "networkDomain"                 "campaign"
[19] "source"                        "medium"
[21] "keyword"                       "isTrueDirect"
[23] "adContent"                     "referralPath"
[25] "adwordsClickInfo.page"         "adwordsClickInfo.slot"
[27] "adwordsClickInfo.gclId"        "adwordsClickInfo.adNetworkType"
[29] "adwordsClickInfo.isVideoAd"    "hits1"
[31] "pageviews"                     "timeOnSite"
[33] "sessionQualityDim"             "newVisits"
[35] "bounces"                       "transactionRevenue"
```

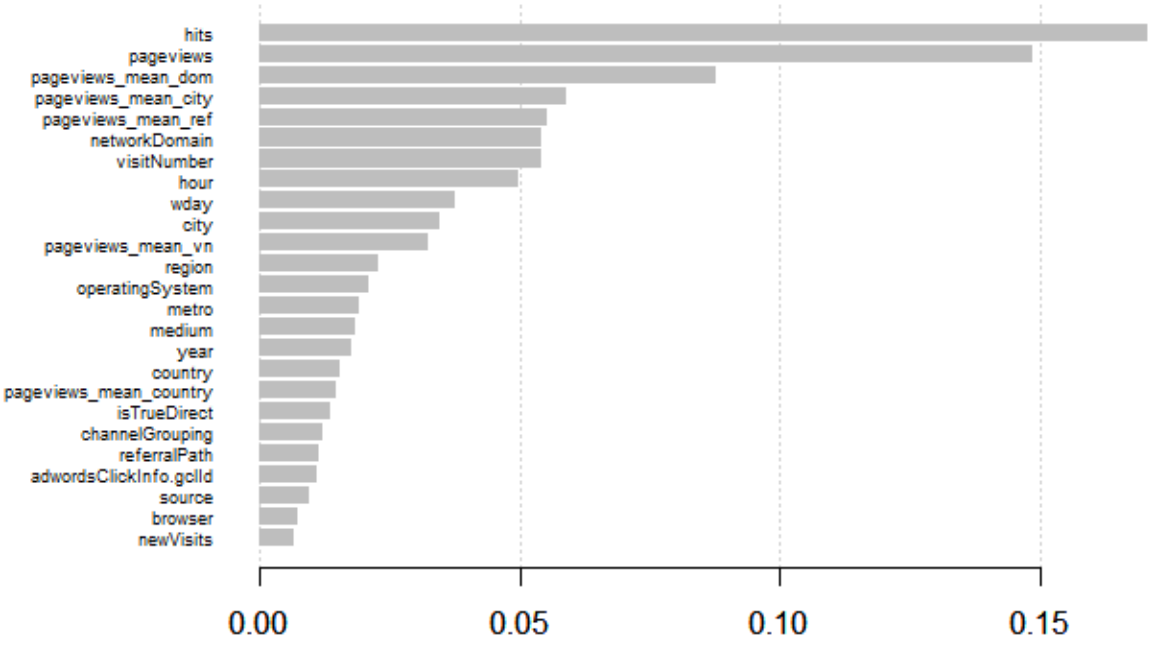
Question 4 : What are the **top three** countries interested in Google swag/products?
Your answer :

Question 5 : On average, which country's customer spends **most**?
Your answer :



Question 6 : What are the **three most** influential indicators for revenue prediction?
Your answer :

Question 7 : What are the **three least** influential indicators for revenue prediction?
Your answer :



```
$ channelGrouping                <fct> Organic Search, Organic Search, Organic Search, Organic ...
$ visitNumber                    <int> 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 3, 1, 5, 2, 1, 1, 1, 1, 11...
$ browser                        <fct> Safari, Chrome, Chrome, Chrome, Safari, Chrome, Chrome, ...
$ operatingSystem                <fct> iOS, Windows, Windows, Windows, Macintosh, Android, Wind...
$ isMobile                       <int> 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0,...
$ deviceCategory                 <fct> mobile, desktop, desktop, desktop, desktop, mobile, desk...
$ continent                      <fct> Americas, Americas, Europe, Asia, Europe, Asia, Americas...
$ subContinent                   <fct> Northern America, Northern America, Southern Europe, Sou...
$ country                        <fct> Canada, Canada, Portugal, India, United Kingdom, Saudi A...
$ region                         <fct> NA, NA, NA, NA, NA, Riyadh Province, New York, NA, New Y...
$ metro                          <fct> NA, NA, NA, NA, NA, NA, New York NY, NA, New York NY, NA...
$ city                           <fct> NA, NA, NA, NA, NA, Riyadh, New York, NA, New York, NA, ...
$ networkDomain                  <fct> NA, NA, vodafone.pt, NA, as9105.com, NA, NA, verizon.net...
$ campaign                       <fct> NA, NA, NA, NA, NA, 1000557 | GA | US | en | Hybrid | GD...
$ source                         <fct> google, google, google, google, google, google, (direct)...
$ medium                         <fct> organic, organic, organic, organic, organic, cpc, NA, or...
$ keyword                        <fct> NA, NA, NA, NA, NA, (User vertical targeting), NA, NA, N...
$ isTrueDirect                   <int> 1, NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, NA, 1, NA, NA,...
$ adContent                      <fct> NA, NA, NA, NA, NA, Google Merchandise Store, NA, NA, NA...
$ referralPath                   <fct> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
$ adwordsClickInfo.page          <fct> NA, NA, NA, NA, NA, 1, NA, NA, NA, NA, NA, NA, NA, 1, NA...
$ adwordsClickInfo.slot          <fct> NA, NA, NA, NA, NA, RHS, NA, NA, NA, NA, NA, NA, NA, Top...
$ adwordsClickInfo.gclId         <fct> NA, NA, NA, NA, NA, CL2-_8Pm9dYCFU9MDQodfdgCIg, NA, NA, ...
$ adwordsClickInfo.adNetworkType <fct> NA, NA, NA, NA, NA, Content, NA, NA, NA, NA, NA, NA, NA,...
$ adwordsClickInfo.isVideoAd     <int> NA, NA, NA, NA, NA, 0, NA, NA, NA, NA, NA, NA, NA, 0, NA...
$ pageviews                      <int> 7, 14, 1, 3, 2, 1, 2, 1, 2, 1, 1, 1, 3, 7, 16, 1, 1, 2, ...
$ newVisits                      <int> NA, 1, 1, 1, 1, NA, NA, 1, 1, 1, NA, 1, NA, NA, 1, 1, 1,...
$ bounces                        <int> NA, NA, 1, NA, NA, 1, NA, 1, NA, 1, 1, 1, NA, NA, NA, 1,...
$ hits                           <int> 7, 18, 1, 3, 2, 1, 2, 1, 2, 1, 1, 1, 3, 7, 21, 1, 1, 2, ...
$ year                           <fct> 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 20...
$ wday                           <fct> 2, 2, 2, 2, 2, 2, 2, 2, 2, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5,...
$ hour                           <fct> 16, 2, 11, 11, 21, 18, 21, 22, 13, 8, 9, 6, 11, 4, 0, 9,...
$ pageviews_mean_vn              <dbl> 4.493438, 3.476790, 3.476790, 3.476790, 3.476790, 4.4934...
$ pageviews_mean_country         <dbl> 5.619792, 5.619792, 3.370370, 2.926108, 2.663194, 3.1764...
$ pageviews_mean_city            <dbl> 7.000000, 14.000000, 1.000000, 3.000000, 2.000000, 1.000...
$ pageviews_mean_dom             <dbl> 7.000000, 14.000000, 5.500000, 3.000000, 1.571429, 1.000...
$ pageviews_mean_ref             <dbl> 7.000000, 14.000000, 1.000000, 3.000000, 2.000000, 1.000...
```

Workshop Summary

What we have experienced/learnt:

- Analyse a case of customer revenue for Google Merchandise Store, using Google Analytics data.
  What's the business value?
- Understand the various features available from Google Analytics data.
  What are some example features?
- Able to conduct exploratory data analysis (EDA) using ***R***
  What are some example EDA charts?
- Able to use predictive models to forecast future customer revenue using ***R***
  What's the predictive models/algorithms we used?
- Able to identify important features which can indicate customer purchases (data insights)
  What are some example data insights?
- Able to build interactive data visualization dashboard to present results & insights using ***Tableau***
  What are ***dimension*** and ***measure*** in Tableau?