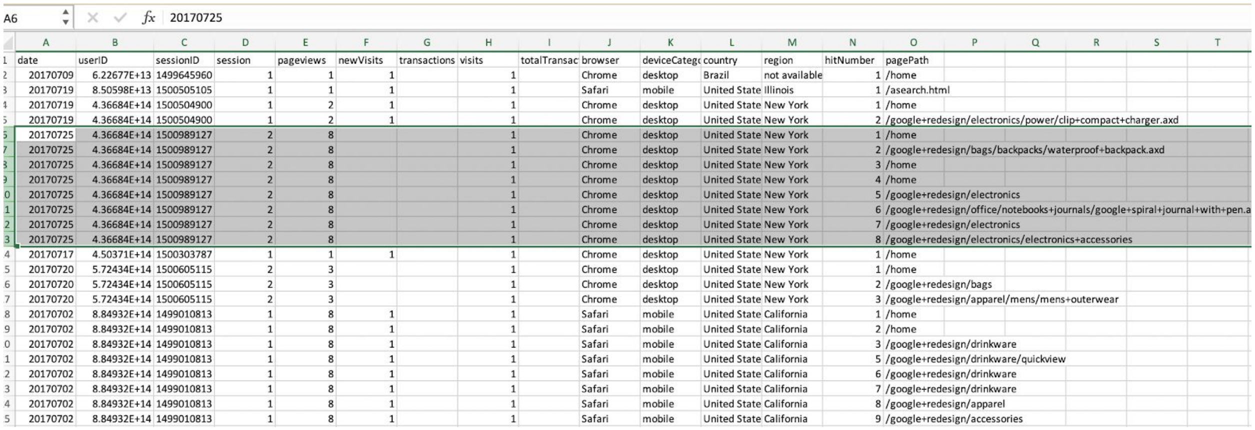
The Data

The sample dataset contains obfuscated Google Analytics 360 data from the Google Merchandise Store, a real ecommerce store that sells Google branded merchandise. The dataset was pulled using Google's BigQuery.

Data dictionary

- **date** - session in YYYYMMDD format
- **userID** - (eg: fullVisitorId) ; The unique visitor ID (also known as client ID).
- **sessionID** - (eg: visitId); Identifier for this session. Only unique to the user. For a completely unique ID, you use a combination of fullVisitorId and visitId.
- **session** - (eg: visitNumber); The session number for this user. If this is the first session, then this is set to 1.
- **pageviews** - Total number of pageviews within the session.
- **newVisits** - Total number of new users in session (for convenience). If this is the first visit, this value is 1, otherwise it is null.
- **transactions** - Total number of ecommerce transactions within the session.
- **visits** - This value is 1 for sessions with interaction events. The value is null if there are no interaction events in the session
- **totalTransactionRevenue** - Total transaction revenue, expressed as the value passed to Analytics multiplied by 10^6 (e.g., 2.40 would be given as 2400000)
- **browser** - browser used (e.g., "Chrome" or "Firefox").
- **deviceCategory** - Type of device (Mobile, Tablet, Desktop).
- **country** - Country from which sessions originated
- **region** - Region from which sessions originate. In the U.S., a region is a state, such as New York.
- **hitNumber** - Sequence of pages that a user looked at within one session. (eg: the sequenced hit number). For the first hit of each session, this is set to 1.
- **pagePath** - URL path of the page.



A6 20170725

date	userID	sessionID	session	pageviews	newVisits	transactions	visits	totalTransac	browser	deviceCatego	country	region	hitNumber	pagePath
20170709	6.22677E+13	1499645960	1	1	1		1		Chrome	desktop	Brazil	not available	1	/home
20170719	8.50598E+13	1500505105	1	1	1		1		Safari	mobile	United State	Illinois	1	/asearch.html
20170719	4.36684E+14	1500504900	1	2	1		1		Chrome	desktop	United State	New York	1	/home
20170719	4.36684E+14	1500504900	1	2	1		1		Chrome	desktop	United State	New York	2	/google+redesign/electronics/power/clip+compact+charger.axd
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	1	/home
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	2	/google+redesign/bags/backpacks/waterproof+backpack.axd
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	3	/home
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	4	/home
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	5	/google+redesign/electronics
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	6	/google+redesign/office/notebooks+journals/google+spiral+journal+with+pen.a
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	7	/google+redesign/electronics
20170725	4.36684E+14	1500989127	2	8			1		Chrome	desktop	United State	New York	8	/google+redesign/electronics/electronics+accessories
20170717	4.50371E+14	1500303787	1	1	1		1		Chrome	desktop	United State	New York	1	/home
20170720	5.72434E+14	1500605115	2	3			1		Chrome	desktop	United State	New York	1	/home
20170720	5.72434E+14	1500605115	2	3			1		Chrome	desktop	United State	New York	2	/google+redesign/bags
20170720	5.72434E+14	1500605115	2	3			1		Chrome	desktop	United State	New York	3	/google+redesign/apparel/mens/mens+outerwear
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	1	/home
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	2	/home
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	3	/google+redesign/drinkware
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	5	/google+redesign/drinkware/quickview
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	6	/google+redesign/drinkware
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	7	/google+redesign/drinkware
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	8	/google+redesign/apparel
20170702	8.84932E+14	1499010813	1	8	1		1		Safari	mobile	United State	California	9	/google+redesign/accessories

Observe that some column (userID, sessionID, etc) values may not be unique. The highlighted observations correspond to a single user visit/session where 8 pages were viewed, but did not result in any transactions. Make sure you perform sufficient analysis to understand what makes an observation unique. **Do not double-count numeric values in calculations, summaries, visualizations, etc.**

Analysis instructions

Perform data analysis as directed below, ensuring to include meaningful explanations of all results for outputs from statistics and visualizations. Organize by using markdown with section headings and text. You are graded on the quality of your visualizations, analysis and report format

- Load the dataset; Inspect its structure (shape/ dimensions), top/bottom rows, summary statistics, etc.
  - Include only variables that make sense in summaries (eg: userID, sessionID, day-of-week is not valuable in such stats)
  - Explain your observations. Sample a couple of observations and explain its meaning.
- Perform data manipulation (as needed) for your analysis and to answer the (5) questions below.
  - Rescale **totalTransactionRevenue** by dividing by 1 million
  - Convert data types as needed to factors, characters, etc;
  - Create new columns as needed (eg: for day-of-week analysis, you will need to create 'dow' variable using lubridate package)
  - Replace NA values as needed. (eg: should NA be replaced with 0?)
- Perform EDA with visualizations (univariate and multivariate analysis), date-time analysis, etc.
  - Do not double-count values in visualizations (use **group_by()** to group variables and **distinct()** to return unique observations)
  - Show your intermediate dataframes in addition to the visualizations

Answer these (5) questions

Make sure to show intermediary data frames/ aggregations/ calculations that help answer the questions. Some hints have been included to assist you but you should apply the tools/ knowledge gained up to this point.

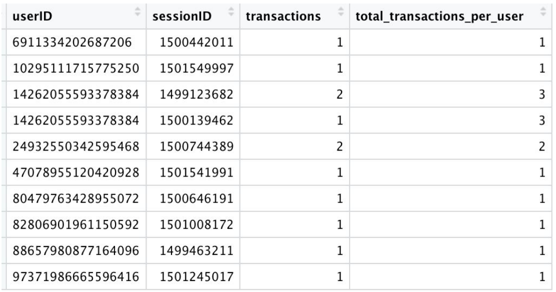
1) What was the average number of product pageviews for users who did make a purchase?

General calculation: SUM(total_pagesviews_per_user) / COUNT(users)

2) What was the average number of product pageviews for users who did not make a purchase?

3) What was the average total transactions per user that made a purchase?

General calculation: SUM (total_transactions_per_user) / COUNT(userID)



userID	sessionID	transactions	total_transactions_per_user
6911334202687206	1500442011	1	1
10295111715775250	1501549997	1	1
14262055593378384	1499123682	2	3
14262055593378384	1500139462	1	3
24932550342595468	1500744389	2	2
47078955120420928	1501541991	1	1
80479763428955072	1500646191	1	1
82806901961150592	1501008172	1	1
88657980877164096	1499463211	1	1
97371986665596416	1501245017	1	1

- Observe that userID in row 3, 4 is for a user who made multiple transactions (2, 1) for a total of 3.
- Observe that the COUNT (userID) is 9 not 10. Only 9 rows correspond to unique userIDs
- Both are calculations you must perform

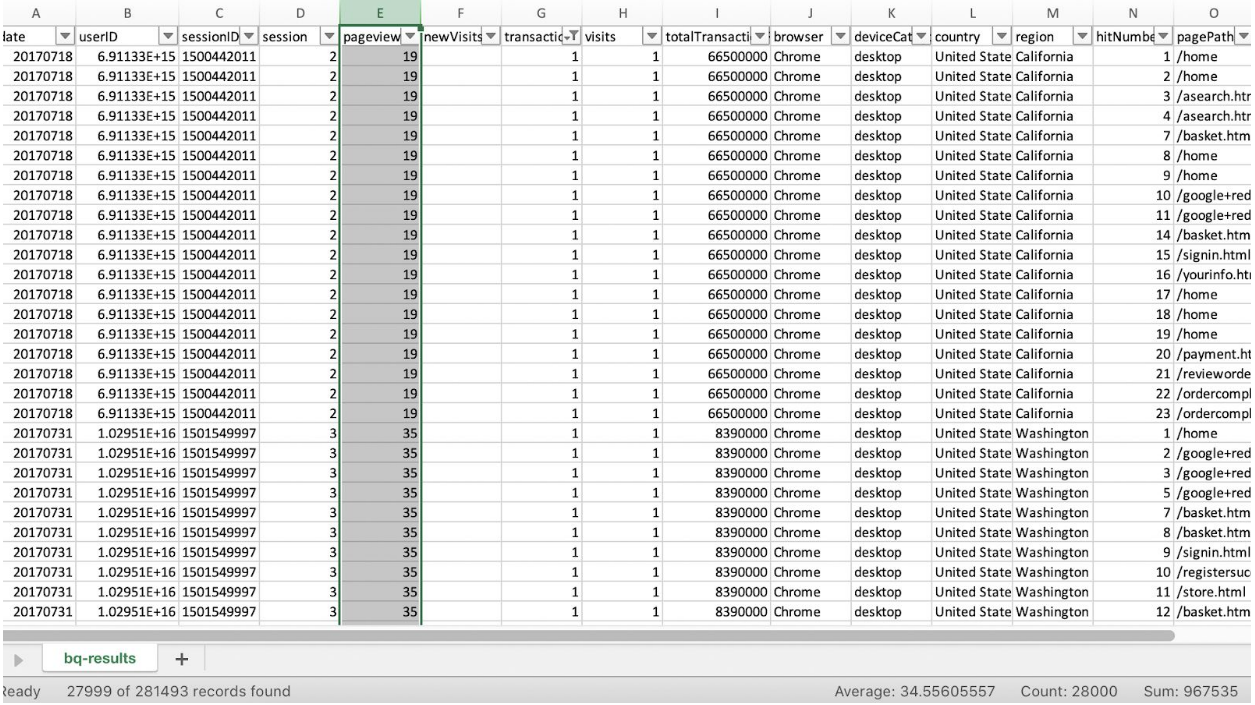
4) What is the average amount of money spent per session? Here per session is the total of 'visits' by user.

General calculation: SUM(total_transactionrevenue_per_user) / SUM(total_visits_per_user)

5) What is the total number of transactions generated per browser type ? Results should be in tabular form that shows the aggregated transactions by browser, including those that resulted in 0 transactions.

Hints:

You can use excel to assist you. For example: To answer #1, you can apply appropriate filters in Excel as shown below. Hovering over the column of interest provides statistics for the variable. In this case the average is **34.55 = 967535/27999**, but that value includes duplicates; pageviews are replicated for each row corresponding to a different url/pagepath. Your calculations should NOT include duplicates when aggregating pageview.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
late	userID	sessionID	session	pageview	newVisits	transactio	visits	totalTransacti	browser	deviceCat	country	region	hitNumbe	pagePath
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	1	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	2	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	3	/asearch.htr
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	4	/asearch.htr
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	7	/basket.htm
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	8	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	9	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	10	/google+red
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	11	/google+red
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	14	/basket.htm
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	15	/signin.html
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	16	/yourinfo.htı
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	17	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	18	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	19	/home
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	20	/payment.ht
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	21	/revieworde
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	22	/ordercompl
20170718	6.91133E+15	1500442011	2	19		1	1	66500000	Chrome	desktop	United State	California	23	/ordercompl
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	1	/home
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	2	/google+red
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	3	/google+red
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	5	/google+red
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	7	/basket.htm
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	8	/basket.htm
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	9	/signin.html
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	10	/registersuc
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	11	/store.html
20170731	1.02951E+16	1501549997	3	35		1	1	8390000	Chrome	desktop	United State	Washington	12	/basket.htm

Ready 27999 of 281493 records found Average: 34.55605557 Count: 28000 Sum: 967535

Extra Credit (+ 5 pts)

Create a model; either

- Linear Regression - continuous outcome variable
- Logistic Regression - binary outcome variable;
  - Ex: predicting conversion. Eg: converted = transactions >= 1 is either True (1) or False (0)
- Example of both types was provided in the sample notebook in your EDA week