SQL Queries + Data Insights for Stakeholders

(These Queries are a part of Jupyter Notebook Which has been Uploaded as well)

Q.Which users spent the Most Money in August?

Which user spent the most money in the month of August?

```
# Pandasql accepts sql lite version of the query. I have mentioned both of the Syntaxes
#         SELECT USER_ID,SUM(TOTAL_SPENT) as Total \
#         from receipts\
#         WHERE extract(year from PURCHASE_DATE) =2022 and\
#         extract(month from PURCHASE_DATE = 8\
#         GROUP BY USER_ID\
#         ORDER BY TOTAL DESC\
#         LIMIT 1"

#Most Money Spent in August 2022
query = "SELECT USER_ID,SUM(TOTAL_SPENT) as Total \
        from receipts\
        WHERE strftime('%Y', PURCHASE_DATE)= '2022' \
        AND strftime('%m', PURCHASE_DATE)= '08'\
        GROUP BY USER_ID\
        ORDER BY TOTAL DESC\
        LIMIT 1"

df = sqldf(query)
df.head()
```

	USER_ID	Total
0	5ffb49a847903912705e9a64	12742.69

Most Money Spent in the month of August irrespective of the Year

```
[35]
query = "SELECT USER_ID,SUM(TOTAL_SPENT) as Total \
        from receipts\
        WHERE strftime('%m', PURCHASE_DATE)= '08'\
        GROUP BY USER_ID\
        ORDER BY TOTAL DESC\
        LIMIT 1"

df = sqldf(query)
df.head()
```

	USER_ID	Total
0	609ab37f7a2e8f2f95ae968f	157739.14

- I Retrieved the User-ID Of the Person with the most Money spent in the Month of August

```
sqlitequery = "SELECT USER_ID,SUM(TOTAL_SPENT) as Total \
      from receipts\
       WHERE strftime('%Y', PURCHASE_DATE)= '2022' \
       AND strftime('%m', PURCHASE_DATE)= '08'\
       GROUP BY USER_ID\
       ORDER BY TOTAL DESC\
       LIMIT 1
```

Q. How many users scanned in each month?

How many users scanned in each month

```
#mysql_query = "SELECT extract(month from DATE_SCANNED) as MONTH_SCANNED,count(USER_ID) as\
#         from receipts\
#         GROUP BY MONTH_SCANNED\
#         ORDER BY TOTAL DESC"


query = "SELECT strftime('%m', DATE_SCANNED) as MONTH_SCANNED,count(USER_ID) as TOTAL_SCANNED \
        from receipts\
        GROUP BY MONTH_SCANNED\
        ORDER BY TOTAL_SCANNED DESC"



df = sqldf(query)
df
```

	MONTH_SCANNED	TOTAL_SCANNED
0	12	8447
1	11	7512
2	10	7305
3	09	6355
4	08	6191
5	07	6058
6	05	5627
7	06	5405
8	04	4882
9	03	4767
10	01	4222
11	02	3830

- The Month of December is when the most number of users Scan via the app

```
sqlitequery = "SELECT strftime('%m', DATE_SCANNED) as
MONTH_SCANNED,count(USER_ID)    as TOTAL_SCANNED \
       from receipts\
       GROUP BY MONTH_SCANNED\
       ORDER BY TOTAL_SCANNED DESC"
```

Q What brand saw the most money spent in June?

What brand saw the most money spent in the month of June

```
# Money Spent in June Throughout
# sql_query = "SELECT BRAND_CODE,SUM(TOTAL_FINAL_PRICE) as Total \
        # from receipt_items\
        # WHERE extract(month from MODIFY_DATE)= '06'\
        # GROUP BY BRAND_CODE\
        # ORDER BY TOTAL DESC\
        # LIMIT 5"

query = "SELECT BRAND_CODE,SUM(TOTAL_FINAL_PRICE) as Total \
        from receipt_items\
        WHERE strftime('%m', MODIFY_DATE)= '06'\
        GROUP BY BRAND_CODE\
        ORDER BY TOTAL DESC\
        LIMIT 5"

df = sqldf(query)
df.head()
```

	BRAND_CODE	Total
0	None	179922.28
1	KIRKLAND SIGNATURE	2610.67
2	GREAT VALUE	1543.84
3	MEMBER'S MARK	819.93
4	KROGER	785.29

```
[57] #Money spent in June 2022
query = "SELECT BRAND_CODE,SUM(TOTAL_FINAL_PRICE) as Total \
        from receipt_items\
        WHERE strftime('%Y', MODIFY_DATE)= '2022'\
        AND strftime('%m', MODIFY_DATE)= '06'\
        GROUP BY BRAND_CODE\
        ORDER BY TOTAL DESC\
        LIMIT 5"

df = sqldf(query)
df.head()
```

	BRAND_CODE	Total
0	None	112145.74
1	KIRKLAND SIGNATURE	1822.17
2	GREAT VALUE	1185.49
3	ANDERSEN	706.00
4	CARDELL	556.98

- Kirkland Signature has spent the most amount of money in June

```
sqlitequery = "SELECT BRAND_CODE,SUM(TOTAL_FINAL_PRICE) as Total \
        from receipt_items\
        WHERE strftime('%m', MODIFY_DATE)= '06'\
        GROUP BY BRAND_CODE\
        ORDER BY TOTAL DESC\
        LIMIT 5"
```

3. Points to share with a Non technical StakeHolder

- After Analyzing the data thoroughly, the following are a few noteworthy points which I feel would be important to share:
    - There are more than 5 columns in the Receipts Data which have more than 90% null values. If these columns are important, we need to think of a way to populate them, otherwise we could do away with these completely
    - Similarly, in the Receipt_items data, Points Earned column has more than 90% values missing
    - Florida, New York, Pennsylvania,Texas and California are the states which contribute to the most users on the App
    - Snacks and Beverages are the top two categories in our Brands Data