Data Analysis

Problem Statement with the data:

Your task is the following:

1. Provide a set of recommendations on how to improve our business or product based on the attached dataset (dataset on car servicing company, company services the customer car upon a online website request a created by customer) .
2. Choose one of the recommendations/insights you uncovered (in #1) and outline an experiment you would like to run to test your suggested product/business recommendation. Please state your hypothesis, describe how you would structure your experiment, list your success metrics and describe the implementation.
3. Let's assume that the experiment you ran (in #2) proved your hypothesis was true. How would you suggest implementing the change on a larger scale? What are some operational challenges you might encounter and how would you mitigate their risk?

Column Names:
orderid: unique identifier of order
parentorderid: unique identifier of order that may contain associated children orders
contactcustomerid: unique identifier of customer
servicecenterid: unique identifier of service center
createdat: timestamp that customer placed the order
pickupdate: timestamp that driver picked up the order
closingdate: timestamp of vehicle return and order close
finalinvoice: final invoice amount to customer
tip: customer tip amount
promocodediscount: discount value of order
grossrevenue: finalinvoice + tip + promocodediscount
netrevenue: revenue after deducting: payment to service centers for work, parts costs, corporate discount promo codes, warranty and returns.

1. Assumptions and data removal steps:

1) If the (Pickup date, closed date is Null) or (created date = pickup date = closing date) and final Invoice amount is zero, I assumed that customer cancelled the order so dropped those observations accordingly, as the orders are cancelled.

2) Removed the observations where closing date is smaller than pickup date (i.e. closing date occurs before pickup date) and pick update is smaller than create date (i.e. pickup occurs before the order is created)

3) Few Observations single order id is associated with multiple service centres and have same final invoice amount as well as net revenue amount. So, I have retained the First observation and removed subsequent observations if a single order id is associated with multiple service centres with same invoice and net revenue amount.

4) If a service centre is associated with more than 20 orders then I assume that the service centre is popular compared to other service centres.

2. New Variables Created

days_to_respond: Days required to respond to an order. It is calculated by subtracting the "order_created_date" and "pickupdate"

mean_days_to_respond: It is calculated for individual service center. For Individual service center, it represents average no of days_to_respond to a order.

daystocompwork: Indicates Days required to complete a work. This is generated by subtracting "pickup date" and "closing date".

mean_days_to_comp_work: It is calculated for individual service center. For Individual service center, It represents the average number of days to complete a work.

total_delivery_time: Days required for work to be completely delivered back to customer is calculated by subtracting the "order_created_date" and "closing date ".

mean_ total_delivery_time: It is calculated for individual service center. For Individual service center, what is the average number of days to deliver a work completely back to customer.

3. Exploratory Data Analysis

Initial Statistical Overview of Numerical Variables:

Based on the given Dataset carDash has 309 service centers and 3977 of customers

	orderid	parentorderid	contactcustomerid	servicecenterid	finalinvoice	tip	promocodediscount	grossrevenue	netrevenue
count	6562.000000	419.000000	6463.000000	4520.000000	6562.000000	6562.000000	6562.000000	6562.000000	6562.000000
mean	2881.164432	2193.565632	1936.368405	195.098009	167.806574	2.110483	6.502560	176.223019	41.383773
std	1668.365779	1465.594956	1166.142990	274.775672	395.303036	7.811407	11.319674	398.669063	177.097279
min	1.000000	40.000000	2.000000	2.000000	0.000000	0.000000	0.000000	-93.010000	-1769.410000
25%	1446.000000	867.500000	939.500000	24.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	2863.500000	2027.000000	1818.000000	46.000000	39.000000	0.000000	0.000000	48.725000	9.010000
75%	4345.750000	3117.000000	2893.000000	312.000000	177.120000	0.000000	10.000000	187.060000	45.000000
max	5764.000000	5705.000000	4187.000000	862.000000	8153.900000	187.010000	75.000000	8153.900000	4288.470000

TakeAways from above View for further analysis:

- count is different for different columns, indicates the data has missing values which we have to deal with them

- gross revenue and net revenue are negative so we have to take a look at those points and determine why they are negative

Insight 1 (**Theme:** Cancelled Orders):

which service center is associated with high proportion of cancelled orders

Based on Assumption 1, I determined whether an order is cancelled or not

1) According to assumption 1 I have created a new column whether the observation is cancelled or not. After that I calculated the proportion of cancelled orders associated with respective service ids. In real time this query will be useful for finding out the service centres that are associated with higher proportion of cancelled orders. Based on that further infestation can be done why the proportion is high

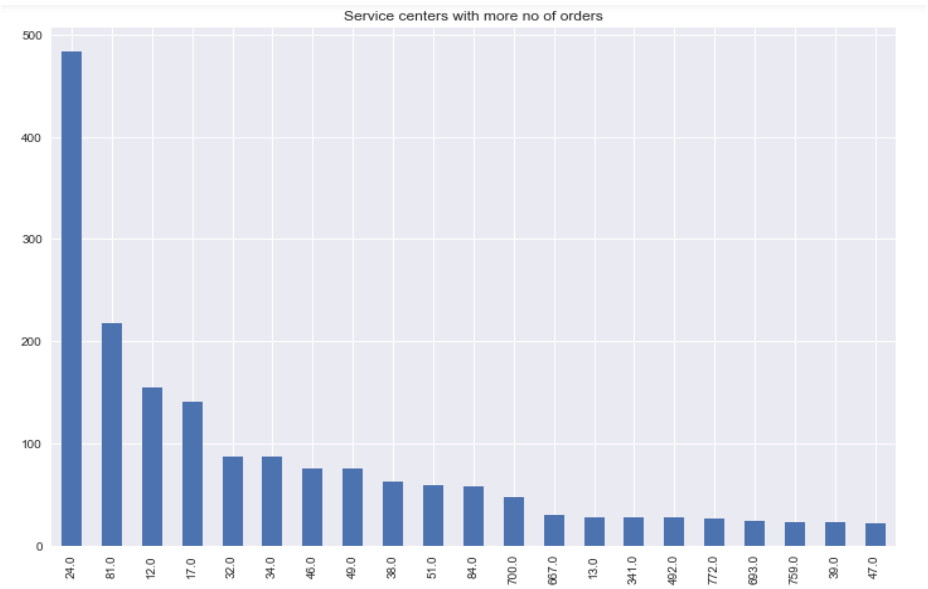
2) Analysis is done for service centres which have at least 10 orders in Total

.

	servicecenterid	Total_orders	no_of_orders_cancelled	proportion_order_cancelled
69	807.0	22	9	0.409091
39	409.0	13	5	0.384615
62	759.0	36	10	0.277778
46	646.0	20	5	0.250000
53	700.0	69	15	0.217391
16	63.0	28	6	0.214286
40	492.0	36	7	0.194444
17	64.0	11	2	0.181818

Insight 2: (Theme: Popular Service Centres)

A service centre is considered popular, if the service centre is associated with **more than 20 orders** (also mentioned in assumptions.



Service center with id 24, followed by 84 have highest no of orders

➡ **Insight 3 (Theme**: Get some free money**)**

Not a strong relation but usually less the total delivery time higher is the tip generated, better reduce the service delivery time to generate higher tip which contributes to net revenue.

Insight 4 (Theme: Being Prepared for future**.)**

We can see start of weekdays (Monday, Tuesday, Wednesday) the count of orders created is higher, whereas the weekend approach the count decreases.

As the weekend approach, Saturday and Sunday people want to hang out with the family or might on a trip with friends, so people will plan to have their car get serviced by weekend and have a good nice weekend trip without any hindrance, leading to request for a service order at the start of the week.

So, if a respective service centre fixes the existing orders by the end of current week and delivering back to the customer might improve the customer satisfaction level.

This is because delivering the service request by weekend frees up the garage space, so the service centre can respond on time to the upcoming orders. Being responsive to a customer service request increases the customer satisfaction.

Insight 5(Theme: Penalizing Bad Performers in the league**)**

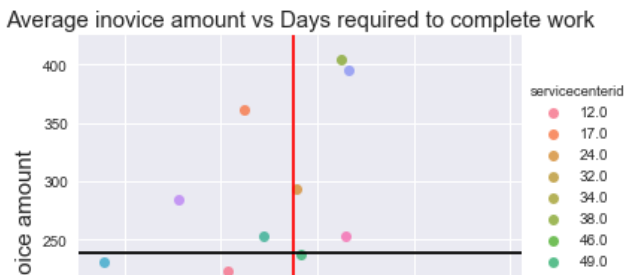
Assumption:

a) complex work is associated with high invoice amount and takes more no of days to resolve the service request. Less complex work is associated with less invoice amount and takes fewer days to resolve the service request.
b) Only popular service centres are considered for this observation (Popular service centers are the service centers associated at least 20 order counts)

Based on the above assumptions There might be few service centres which are associated with very less average invoice amount but take higher average no of days to complete the work.

It might be possible that these centres take more time for the same job which can be done in few days by other service centres, car dash can give warning to those service centres and provide better experience to customers by giving them best choice of service centres to choose from.

Below is the image which represent one few of those service centre



Example of Bad performer (759) in the league the average invoice amount associated with that

Vertical red line: Overall average no of days to deliver car back to customer by all service centres.

Horizontal Black line: Overall average invoice amount associated with all service centre.

Insight 6(**Theme:** Don't concentrate only on the existing work too much, allocate some time in responding to others as well):

X- Axis represents: Service centre ID

Right Y axis: Average No of Days per service centre to complete the service request (Closing date- pick up date), red line represents this

Left y Axis: Average No of Days per service centre to complete the service request (pickup date- create date), red line represents this, vertical bar plots represent this

Legend

(Average No of Days per service centre to complete the service request (Closing date- pick up date i.e)

(Average No of Days per service centre to complete the service request (pickup date- create date), red line represents this, vertical bar plots represent this)

Dual Axis Chart

It might be possible these yellow set of Service centres have less no of workers who can complete the work fast (represented by red line), but as the no of workers are less, they respond late to a new service request due this their count of orders might be low represented by width of bar plots). Because customer prefers a service centre, which responds back to them on time, especially when they need to service a vehicle. (this can be observed with black boxed service centres)

Yellow set service centres: 759, 667, 772, 341, 700, 49, 51

Black set Service centres might be having more no of workers, out of which few completes the existing work (not too fast, or not too slow, represented by red line), other few workers respond to a new service requests providing better customer satisfaction. Due to this higher no of orders might be associated with black set of service centres compared to yellow set service centres. so, customers might be attracted to these service centres as they have balance between response rate and work completion rate.

Red set service centres: 17, 46, 81, 34, 32, 24, 12, 38

ons/insi
run t
ypothe
nd desc

I choose **Insight 6** (**Theme:** Don't concentrate only on the existing work too much, allocate some time in responding to others as well):

Hypothesis: Having few workers and Concentrating on completing the work associated with ongoing service requests by neglecting providing response to new service requests might affect the customer satisfaction level. This might be the reason for incurring few no of service requests to that service centre.

Experiment: Surveying the People by asking them on what level they prefer faster response compared to

Implementation: (Including a survey after delivering the service)

Structuring the experiment:

3. Let's assume that the experiment you ran (in #2) proved your hypothesis was true. How would you suggest implementing the change on a larger scale? What are some operational challenges you might encounter and how would you mitigate their risk?

   Implementing change, contact service centres and ask them to increase the no of workers as well garage space

   Operational challenges

Problem Statement

Please complete the following 2 SQL questions writing out your queries longhand. As you are not working in a live database environment, please make and show assumptions. The purpose is to understand your thought process around data querying and manipulation. The table schema and column definitions are provided below.

Weighted average retention looks at all new customers (they are new in month 0 and so month 0 retention is 100%) and looks at their behavior in month 1. For new customers in October 2017, month 1 is November 2017. For new customers in December 2017, month 1 is January 2018. We take a weighted average of all the 1-month-after-purchase retention numbers to get a 1-month retention calculation.

a. Please write a query to establish monthly cohorts based on month joined.
b. Please write a query or series of queries to calculate weighted average retention of new customers by months after first purchase (1,2,etc.)

Relevant tables and fields:
Format: table_name.field_name
Example: delivery.actual_delivery_time

Tables

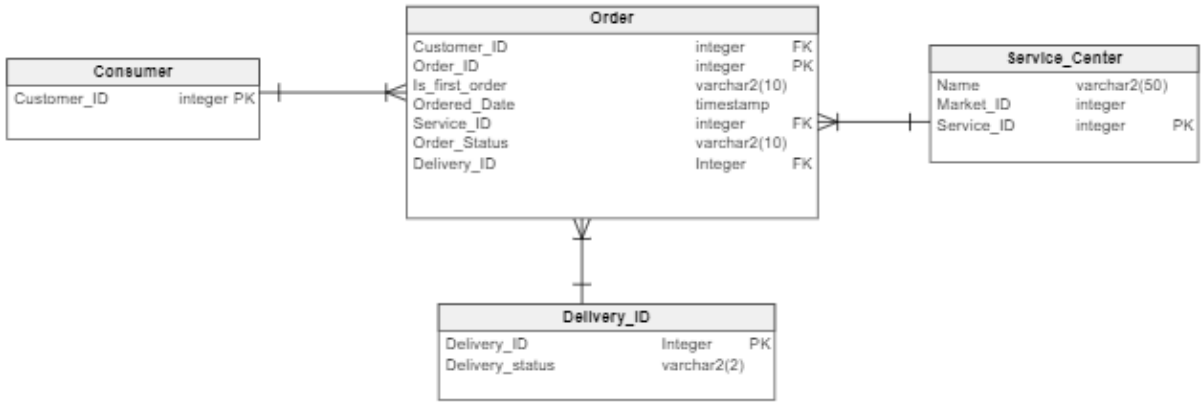
consumer
consumer id: unique ID of consumer

order
id: unique ID given to each order
creator_id: consumer ID of order creator
delivery_id: delivery ID associated with order
is_first_ordercart: designation of order from a first-time customer

service center
id: unique ID of service center associated with delivery
name: service center name
market_id: ID number of market in which service center is located

Analysis

Based on the given tables in the description, I have drawn the following Normalized Entity-Relationship diagram based on few assumptions. I have included only important columns, that are required to track an order associated with a particular customer



Assumptions of ER diagram:

1) A customer is associated with at least order.

2) One service center will be associated with at least order or to many orders
3) A delivery_ID can consist of one to many orders (As sometimes customers order only one item or sometimes more items to same delivery address,)

Based on my understanding, assumptions and assignment queries, I found that only customer_id and his transaction dates are required to solve the query 1.

As I am not provided with live database environment I created a table in SQL developer with customer_id and his transaction dates as columns in the table. The data I used is a super market dataset, which I got from online.

Assumption for solving queries:

A customer is considered as a new user to a particular month on the basis of his first order month. If customer orders his first service request or item in September, then the customer is considered as new user for September

Query 1:

Step 1: Created the required table

```
CREATE TABLE intern
  (
     cust_id          VARCHAR2(20),
     transaction_date DATE
  );
```

Step 2: Loaded the data into database from excel (cust_id, transaction_date columns)

	CUST_ID	TRANSACTION_DATE
1	PO-19195	04-01-14
2	PO-19195	04-01-14
3	PO-19195	04-01-14
4	MB-18085	05-01-14
5	LS-17230	06-01-14
6	JO-15145	06-01-14
7	ME-17320	06-01-14
8	ME-17320	06-01-14
9	ME-17320	06-01-14
10	ME-17320	06-01-14

Step 3: Created a Visit log view. This visit log consists of history or activity of a customer. In what month a customer is making an order or does some activity. (For example, the customer AA-10315 made orders in 2, 8, 3, 9 months)

```
CREATE VIEW visit_log
AS
  SELECT cust_id,
         Extract(month FROM transaction_date) - 1 AS visit_month
  FROM   dataset
  GROUP  BY cust_id,
```

```
        Extract(month FROM transaction_date) - 1
ORDER  BY cust_id,
        Extract(month FROM transaction_date) - 1;
```

	CUST_ID	VISIT_MONTH
1	AA-10315	2
2	AA-10315	8
3	AA-10375	3
4	AA-10375	9
5	AA-10480	4
6	AA-10645	5
7	AA-10645	11
8	AB-10015	1
9	AB-10015	2
10	AB-10105	11

Step 4: Using **visit_log** created **cust_first_visit** view. This table indicates customer id and the associated month of his first order.

```
CREATE VIEW cust_first_visit
AS
  SELECT cust_id,
         Min(visit_month) AS first_month
  FROM   visit_log
  GROUP  BY cust_id;
```

	CUST_ID	FIRST_MONTH
1	AH-10585	9
2	RA-19915	5
3	MS-17770	6
4	DP-13000	10
5	ML-17410	10
6	MM-18280	0
7	SD-20485	0
8	CK-12325	2
9	JF-15295	2

Step 5: Created a view for calculating count of new users in individual months (Monthly cohort size).

```
CREATE VIEW monthly_new_users
AS
  SELECT first_month              month_number,
         Count(DISTINCT cust_id) AS new_users
  FROM   cust_first_visit
  GROUP  BY first_month;
```

	MONTH_NUMBER	NEW_USERS
1	1	24
2	6	44
3	11	49
4	2	65
5	4	56
6	5	48
7	8	68
8	7	49
9	3	56
10	10	63

Step 6: view to find out whether a new user retained in subsequent months (new user in, Month 1, returned in both Month 2 and 3 or only in Month 3). **(**For example, customer AA-10315 in below table has first order in month of February (w.r.t to him month 0, again he ordered in July (it will be his month 6)

```
CREATE VIEW user_activities
AS
  SELECT vl.cust_id,
         vl.visit_month - cust_first_visit.first_month MONTH_NUMBER
  FROM   visit_log vl
         left join cust_first_visit
                ON vL.cust_id = cust_first_visit.cust_id
  GROUP  BY vl.cust_id,
            vl.visit_month - cust_first_visit.first_month
  ORDER  BY vl.cust_id,
            vl.visit_month - cust_first_visit.first_month;
```

	CUST_ID	MONTH_NUMBER
1	AA-10315	0
2	AA-10315	6
3	AA-10375	0
4	AA-10375	6
5	AA-10480	0
6	AA-10645	0
7	AA-10645	6
8	AB-10015	0
9	AB-10015	1
10	AB-10105	0

Step 7: Claculation of number of users retained or returned in a subsequest month w.r.t to the total number of users in the starting month

```
CREATE VIEW retention_table
AS
  SELECT C.first_month         cohort_month,
         A.month_number,
         Count(C.first_month) AS num_users
  FROM   user_activities A
         left join cust_first_visit C
                ON A.cust_id = C.cust_id
  GROUP  BY C.first_month,
            A.month_number
  ORDER  BY C.first_month,
            A.month_number;
```

```
SELECT *
FROM   retention_table;
```

	COHORT_MONTH	MONTH_NUMBER	TOTAL_USERS	RETAINED_USERS	PERCENTAGE
1	0	0	31	31	100
2	0	1	31	3	9.68
3	0	3	31	2	6.45
4	0	4	31	2	6.45
5	0	6	31	2	6.45
6	0	7	31	4	12.9
7	0	8	31	5	16.13
8	0	9	31	3	9.68
9	0	10	31	6	19.35
10	0	11	31	5	16.13
11	1	0	24	24	100
12	1	1	24	4	16.67
13	1	2	24	2	8.33
14	1	3	24	1	4.17
15	1	5	24	2	8.33
16	1	6	24	2	8.33
17	1	7	24	3	12.5
18	1	8	24	3	12.5
19	1	9	24	4	16.67
20	1	10	24	4	16.67

```
                                  total_users,
                                  retained_users,
ers,                              2) PERCENTAGE

ch_number
```

For better understanding See from row 11 and read below description

Total users column indicates number of new customers that have been created in each cohort month. Moving right down the percentage column for cohort_ month (1 i.e February, i.e at row 11), the percentage of those new cohort_month (1) customers retained back in the months (February, March, April.....) are indicated by month_number.

Ex: Total No of new customer in February is 24 and percentage of retention of those customer in march is 16.67, April is 8.33.....