COURSE: EDA, SQL & BIG DATA

SQL MINI PROJECT: E-commerce analysis

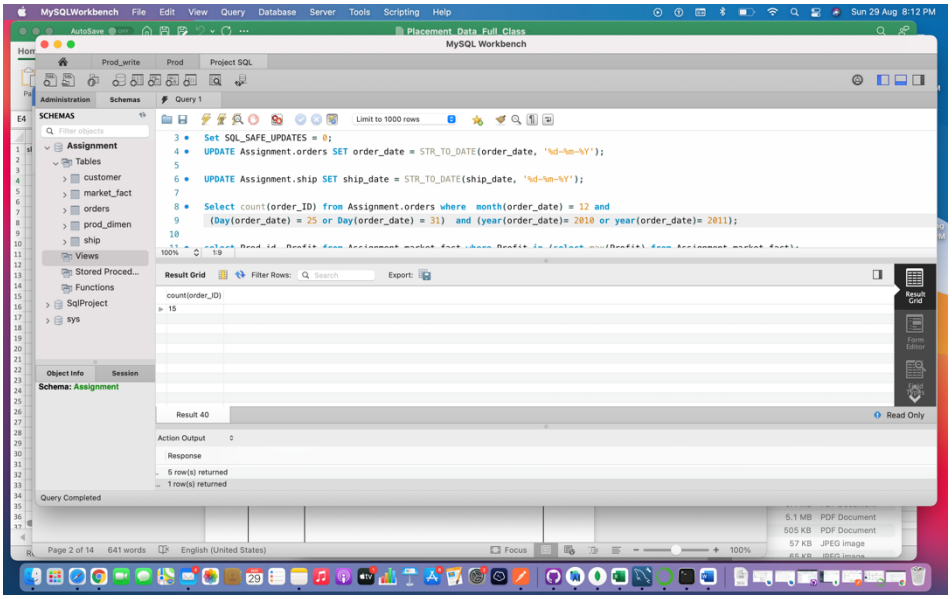
EDA MINI PROJECT: Placement Analysis

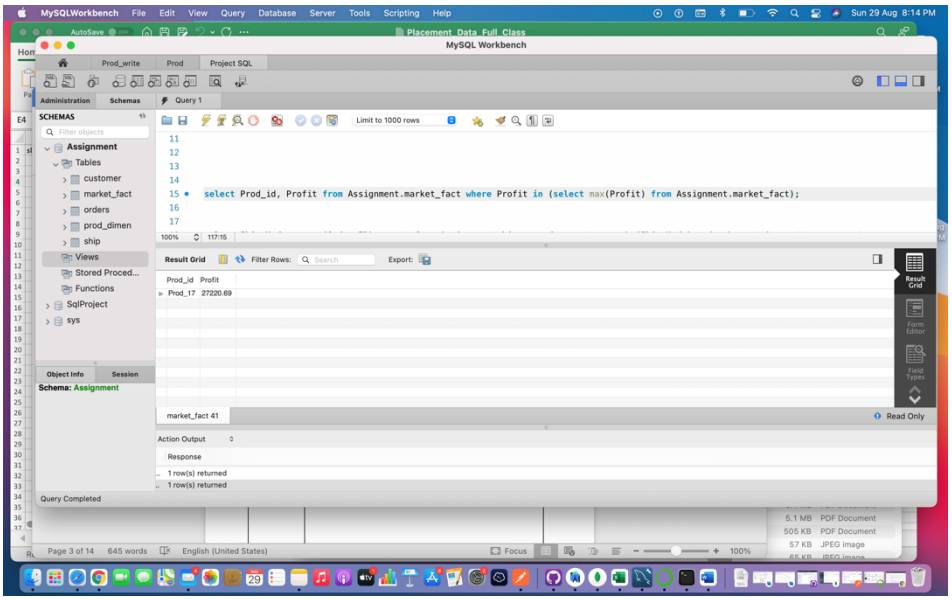
Instructions for the submission:

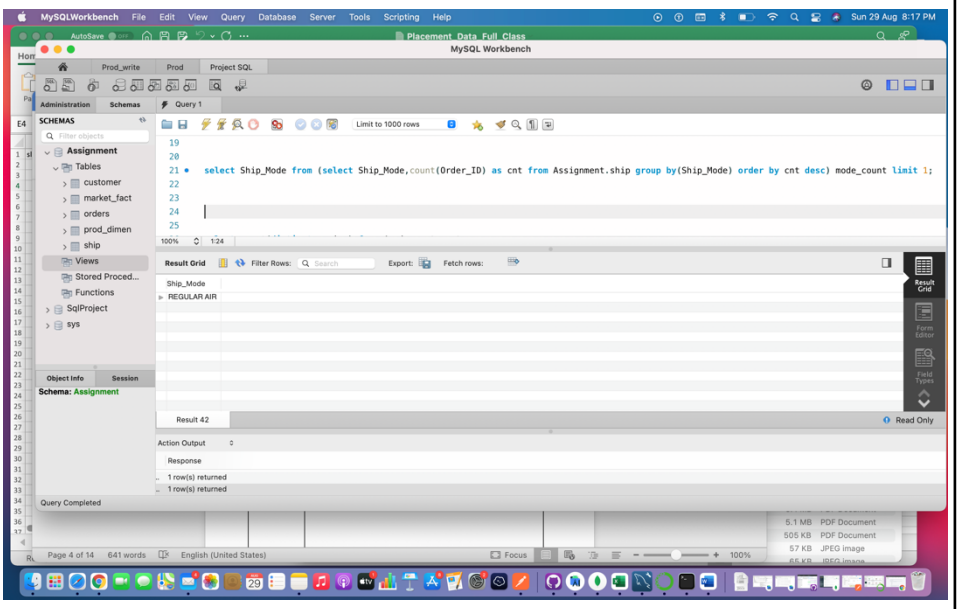
- Please maintain the following: Font - Times New Roman, Font Size - 12, Line Spacing - 1.5
- Wherever mentioned, please ensure that the relevant screenshots of your working from the SQL workbench/Python notebook/Excel Sheet

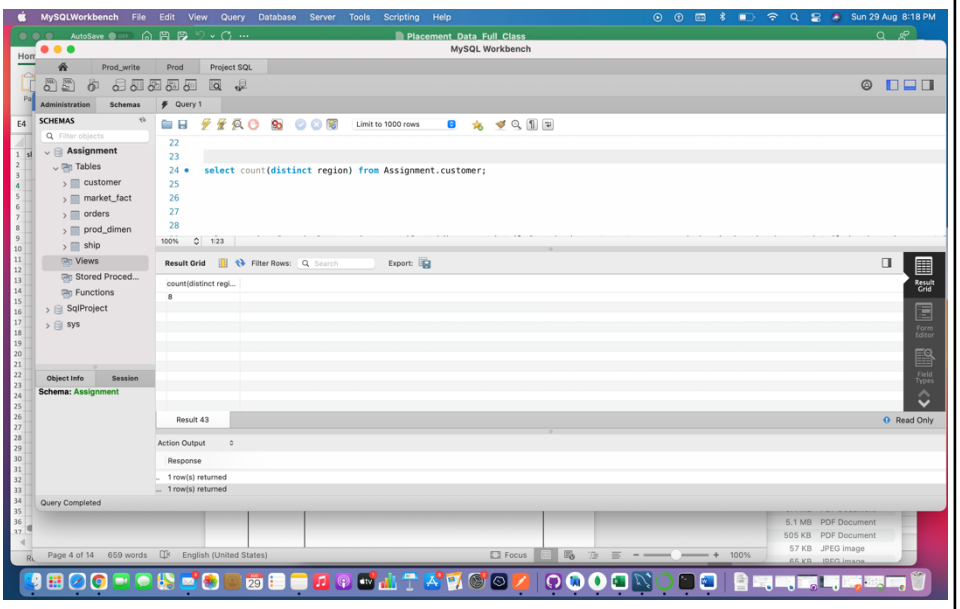
PART I: SQL

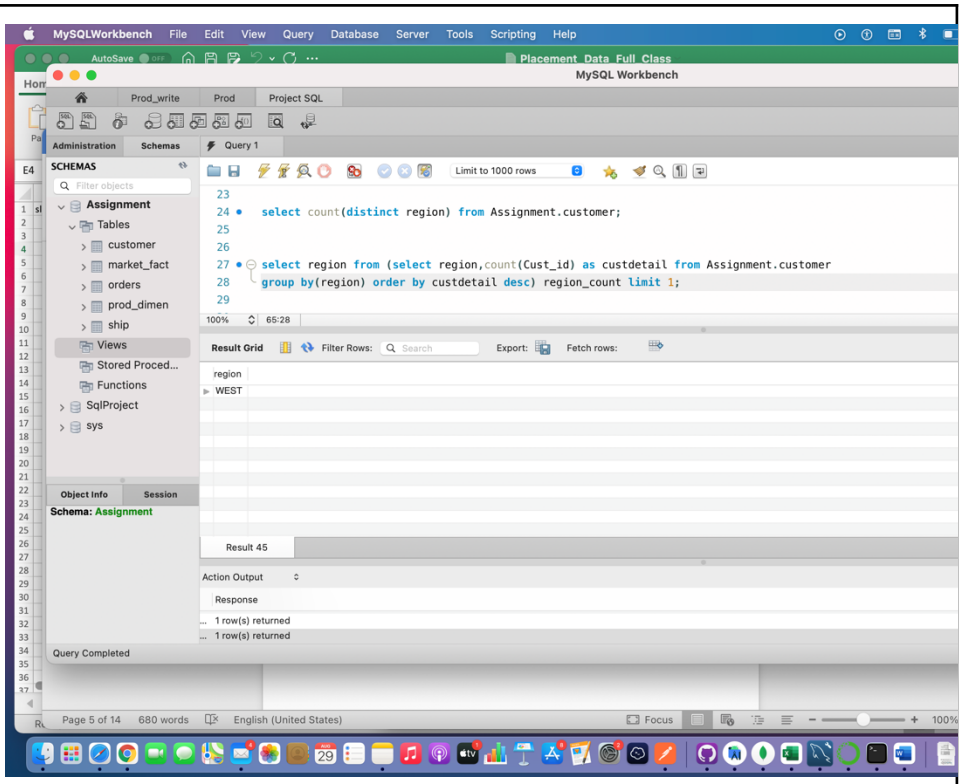
S.No.	Assessment Area	Submission Format	Marks
1	SQL	SQL Query + Screenshot of output	2 marks
2		SQL Query + Screenshot of output	2 marks
3		SQL Query + Screenshot of output	2 marks
4		SQL Query + Screenshot of output	2 marks
5		SQL Query + Screenshot of output	2 marks
6		SQL Query + Screenshot of output	2 marks
7		SQL Query + Screenshot of output	4 marks
8		SQL Query + Screenshot of output	4 marks
Project Part I Maximum Marks			**20 marks**

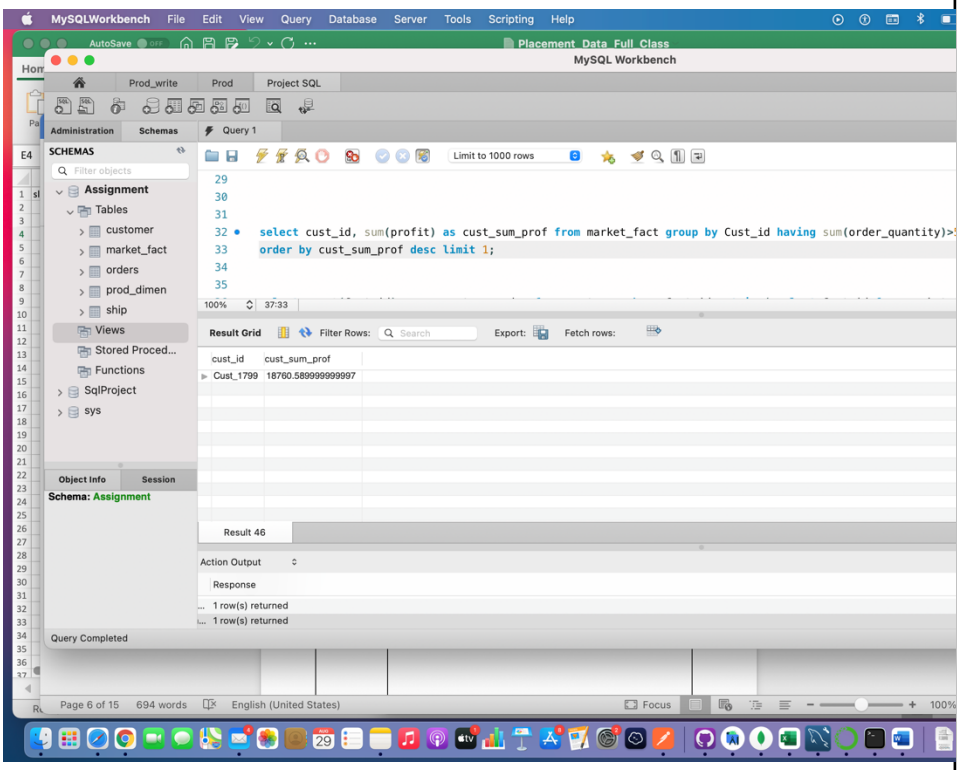
Question 1	Find the number of orders that were made on the most…
Marks	2 marks
SQL Query	Select count(order_ID) from Assignment.orders where month(order_date) = 12 and (Day(order_date) = 25 or Day(order_date) = 31) and (year(order_date)= 2010 or year(order_date)= 2011);
Screenshot of query and output	

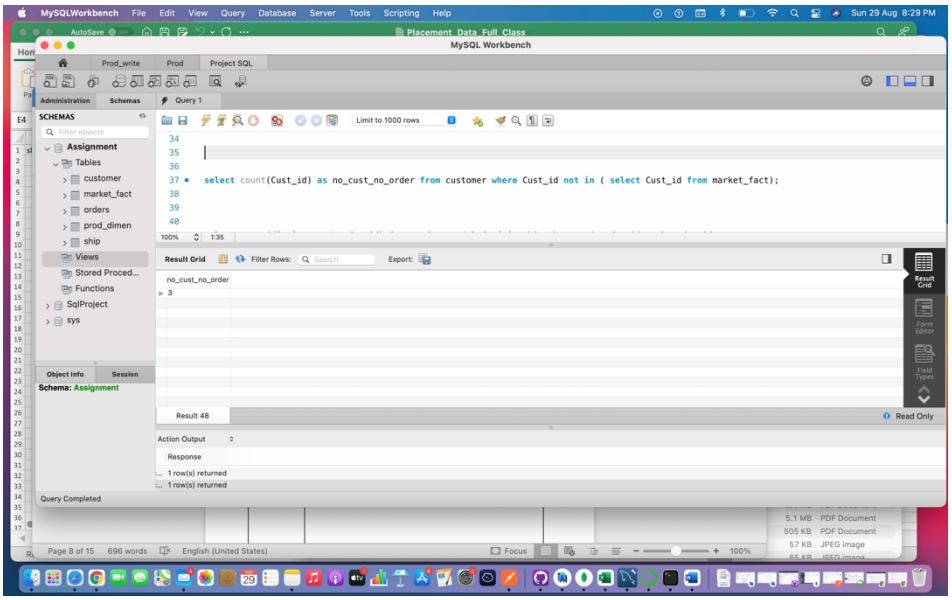
Question 2	You want to focus your marketing and sales resources…
Marks	2 marks
SQL Query	select Prod_id, Profit from Assignment.market_fact where Profit in (select max(Profit) from Assignment.market_fact);
Screenshot of query and output	

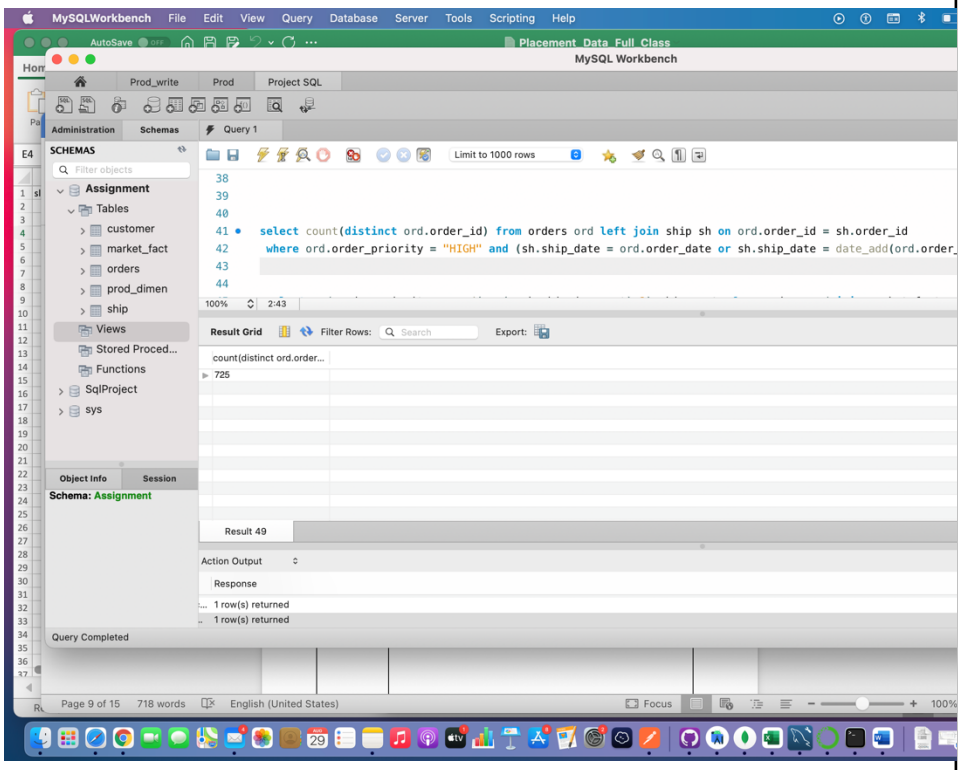
Question 3	Which is the most frequent mode of shipment…
Marks	2 marks
SQL Query	select Ship_Mode from (select Ship_Mode,count(Order_ID) as cnt from Assignment.ship group by(Ship_Mode) order by cnt desc) mode_count limit 1;
Screenshot of query and output	*Paste your screenshots here* 

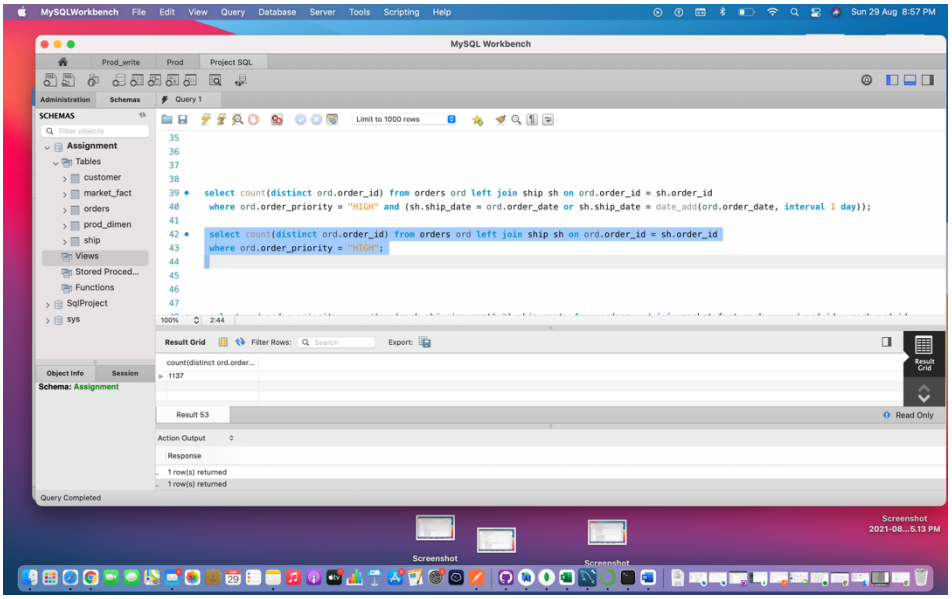
Question 4	What is the number of unique regions that the company…
Marks	2 marks
SQL Query	a) select count(distinct region) from Assignment.customer; b) select region from (select region,count(Cust_id) as custdetail from Assignment.customer group by(region) order by custdetail desc) region_count limit 1;
Screenshot of query and output	*Paste your screenshots here* 


MySQLWorkbench File Edit View Query Database Server Tools Scripting Help
Placement_Data_Full_Class
MySQL Workbench
Prod_write
Prod
Project SQL
Administration
Schemas
Query 1
SCHEMAS
Filter objects
Assignment
Tables
customer
market_fact
orders
prod_dimen
ship
Views
Stored Proced...
Functions
SqlProject
sys
Limit to 1000 rows
23
24 select count(distinct region) from Assignment.customer;
25
26
27 select region from (select region,count(Cust_id) as custdetail from Assignment.customer
28 group by(region) order by custdetail desc) region_count limit 1;
29
100%
65:28
Result Grid
Filter Rows:
Search
Export:
Fetch rows:
region
WEST
Result 45
Object Info
Session
Schema: Assignment
Action Output
Response
1 row(s) returned
1 row(s) returned
Query Completed
Page 5 of 14
680 words
English (United States)
Focus
100%

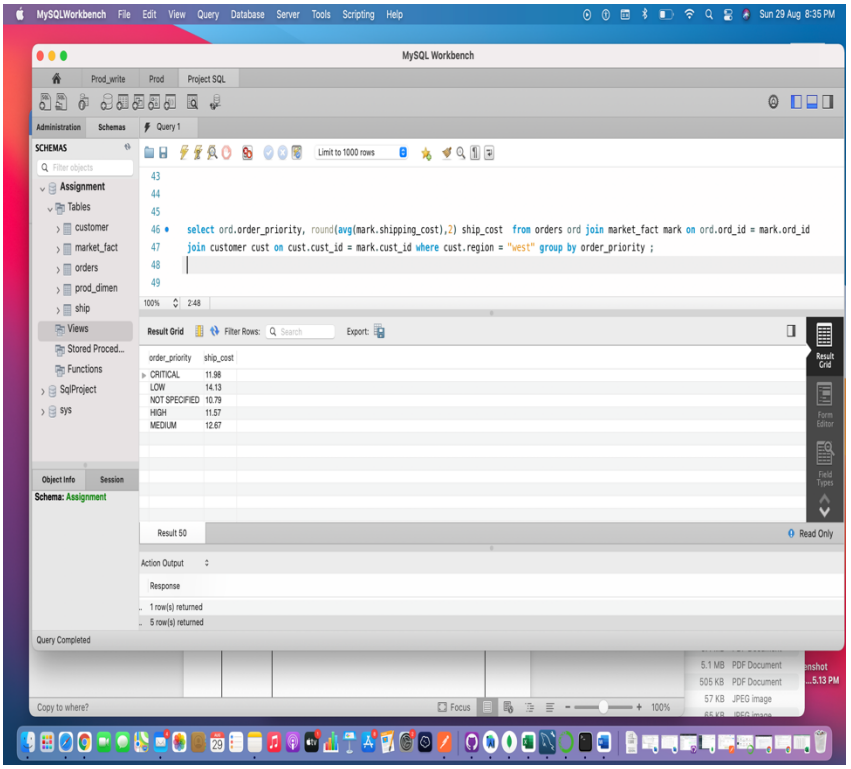
Question 5	To improve your customer retention and give them…
Marks	2 marks
SQL Query	select cust_id, sum(profit) as cust_sum_prof from market_fact group by Cust_id having sum(order_quantity)>500 order by cust_sum_prof desc limit 1;
Screenshot of query and output	

Question 6	Find the number of customers who have not made any…
Marks	2 marks
SQL Query	select count(Cust_id) as no_cust_no_order from customer where Cust_id not in (select Cust_id from market_fact);
Screenshot of query and output	

Question 7	According to the e-commerce company, a High priority order…
Marks	4 marks
SQL Query	a) select count(distinct ord.order_id) from orders ord left join ship sh on ord.order_id = sh.order_id where ord.order_priority = "HIGH" and (sh.ship_date = ord.order_date or sh.ship_date = date_add(ord.order_date, interval 1 day)); b) select count(distinct ord.order_id) from orders ord left join ship sh on ord.order_id = sh.order_id where ord.order_priority = "HIGH";
Screenshot of query and output	

	 The claim made by the company is false because total high priority orders are 1137 but only 735 high priority orders are delivered on the same date or next date of ordered date.

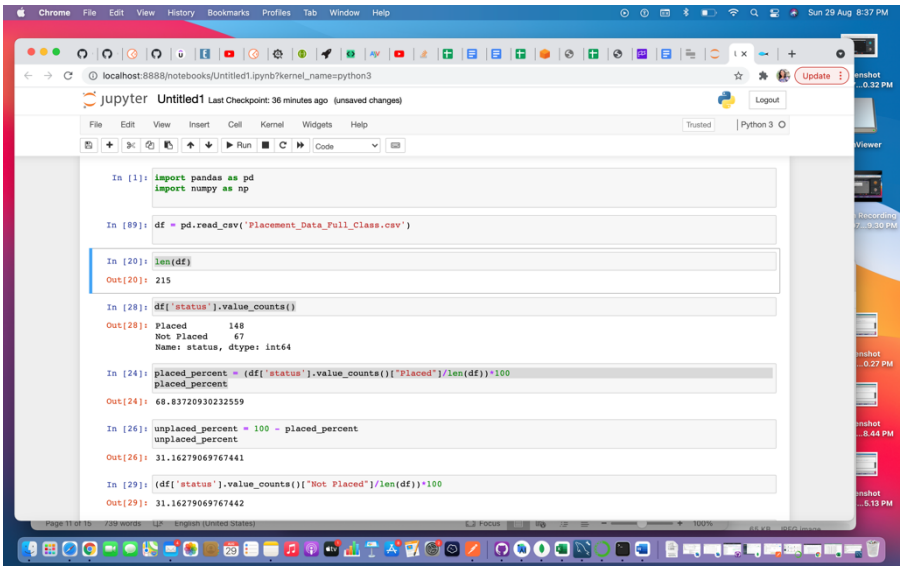
Question 8	Compare the average shipping cost for various “order_priority”…
Marks	4 marks
SQL Query	select ord.order_priority, round(avg(mark.shipping_cost),2) ship_cost from orders ord join market_fact mark on ord.ord_id = mark.ord_id join customer cust on cust.cust_id = mark.cust_id where cust.region = "west" group by order_priority ;

Screenshot of query and output	

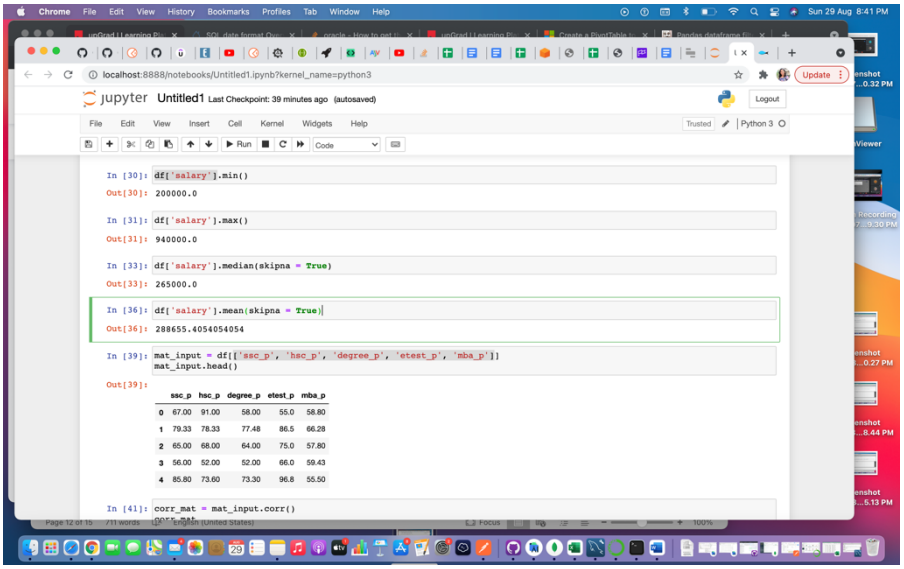
PART II: EDA

S.No.	Assessment Area	Submission Format	Marks
1	EDA	Steps taken/Python code + Screenshot of output	3marks
2		Steps taken/Python code + Screenshot of output	3 marks
3		Steps taken/Python code + Screenshot of output	3 marks

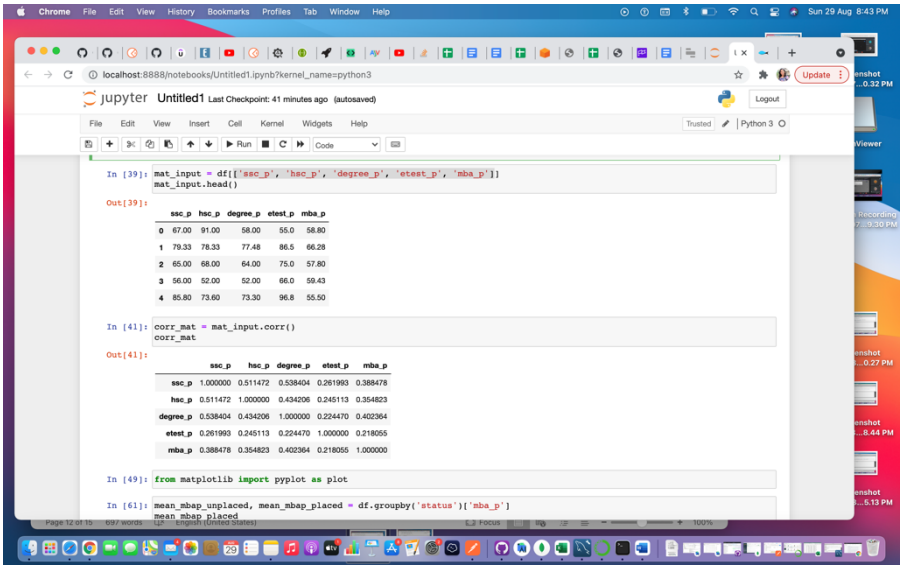
4		Steps taken/Python code + Screenshot of output	3 marks
5		Steps taken/Python code + Screenshot of output	8 marks
Project Part II Maximum Marks			**20 marks**

Question 1	What is the percentage of people placed …..
Marks	3 Marks
Solution	Steps taken :- 1) Calculate the length of dataframe i.e. total no. of students. 2) Used .value_counts() to find out no. of "Placed" and "Not placed" students. 3) Calculated the percentage for the above two by status type diving by total no. of students and multiplying by 100. 

Question 2	Find the minimum, maximum, median and average…..
Marks	3 Marks

Solution	Steps taken:- 1) Use df['salary'].min() to calculate minimum salary. 2) Use df['salary'].max() to calculate maximum salary. 3) Similary used median and mean function to calculate median and average salary offered to students Yes, the average is a good measure of central tendency of data. 

Question 3	Analyse the correlation between the following ……..

Marks	3 Marks
Solution	Steps taken :- 1) Extracted the required columns from the dataframe. 2) Used .corr() to calculate the correlation matrix between all the pairs of columns. My comment: The correlation between these tests is not very prominent as they lie between 20% to 55% only. Highest Correlation: degree_p and ssc_p Lowest Correlation: mba_p and etest_p 

```
In [39]: mat_input = df[['ssc_p', 'hsc_p', 'degree_p', 'etest_p', 'mba_p']]
         mat_input.head()
```

Out[39]:

	ssc_p	hsc_p	degree_p	etest_p	mba_p
0	67.00	91.00	58.00	55.0	58.80
1	79.33	78.33	77.48	86.5	66.28
2	65.00	68.00	64.00	75.0	57.80
3	56.00	52.00	52.00	66.0	59.43
4	85.80	73.60	73.30	96.8	55.50

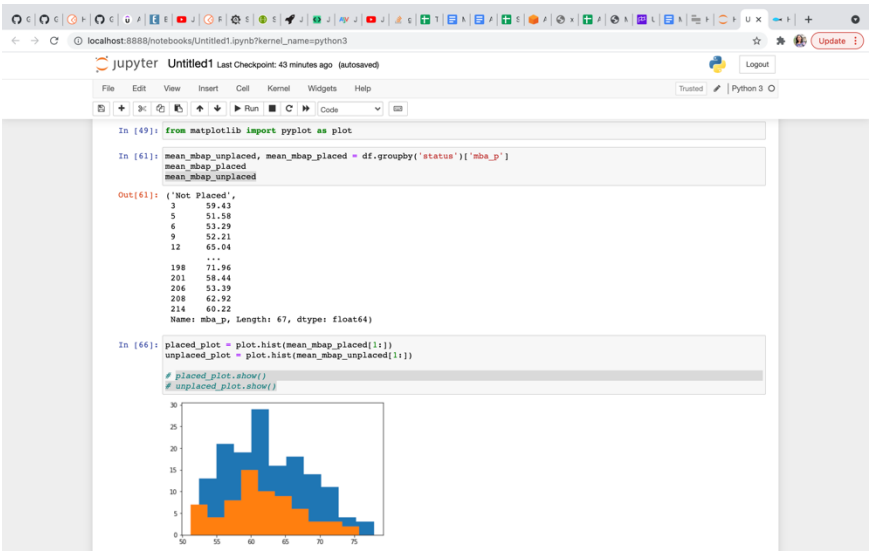
```
In [41]: corr_mat = mat_input.corr()
         corr_mat
```

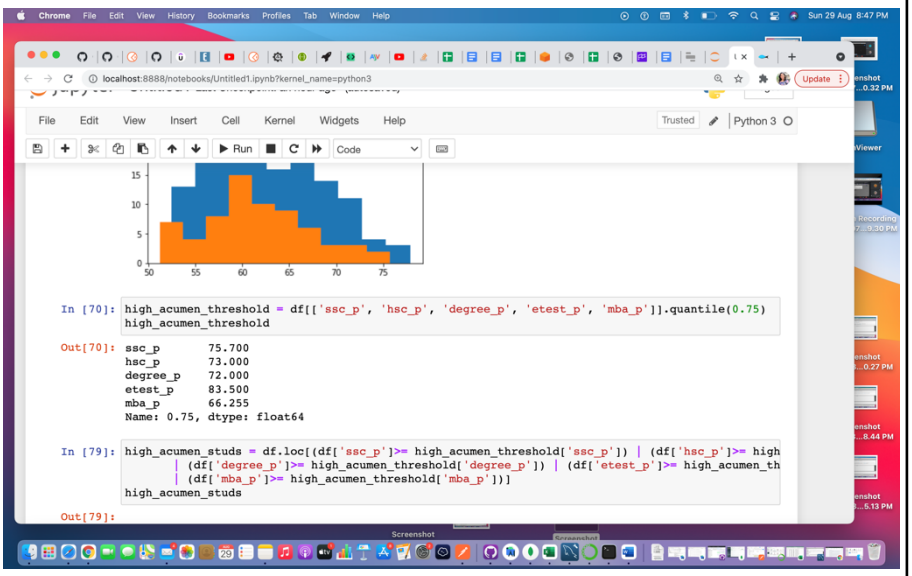
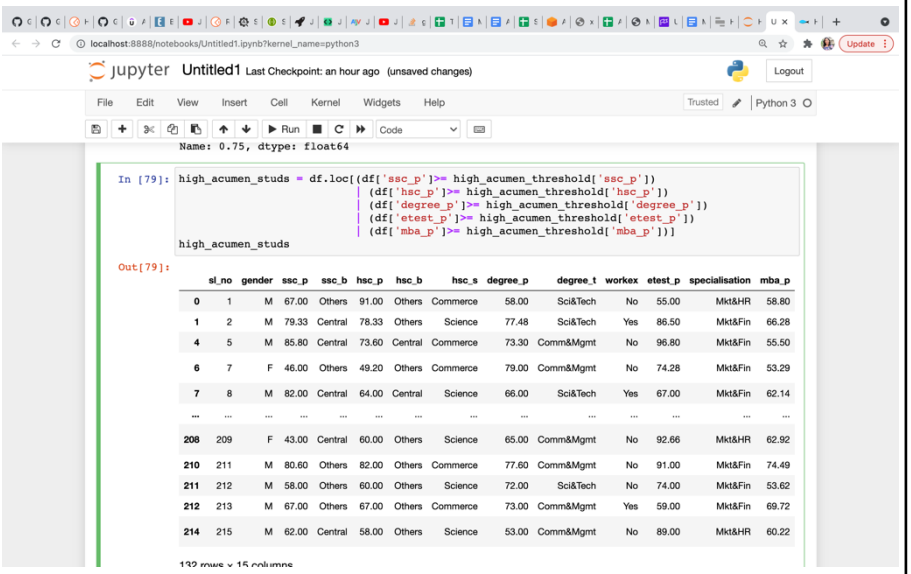
Out[41]:

	ssc_p	hsc_p	degree_p	etest_p	mba_p
ssc_p	1.000000	0.511472	0.538404	0.261993	0.388478
hsc_p	0.511472	1.000000	0.434206	0.245113	0.354823
degree_p	0.538404	0.434206	1.000000	0.224470	0.402364
etest_p	0.261993	0.245113	0.224470	1.000000	0.218055
mba_p	0.388478	0.354823	0.402364	0.218055	1.000000

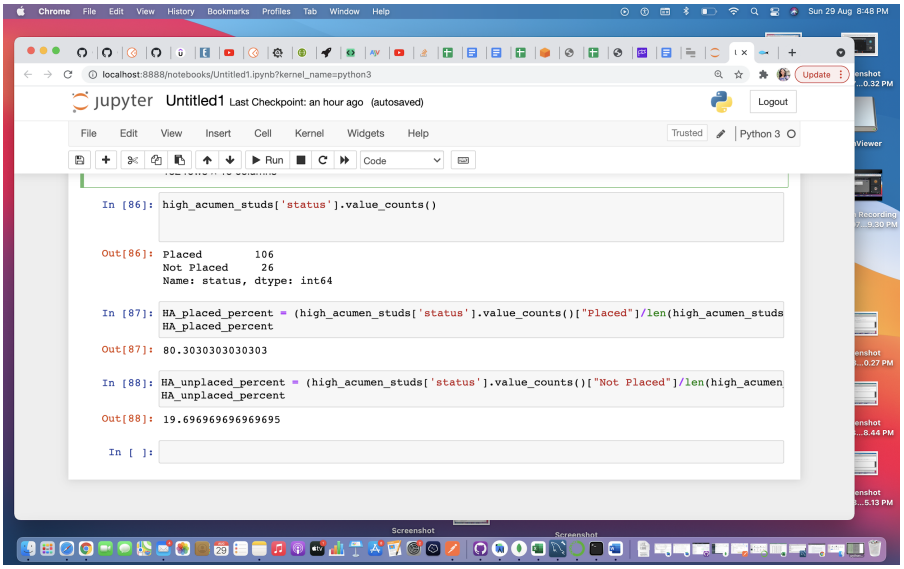
```
In [49]: from matplotlib import pyplot as plot

In [61]: mean_mbap_unplaced, mean_mbap_placed = df.groupby('status')['mba_p']
         mean_mbap_placed
```

Question 4	You want to compare the MBA_P score……
Marks	3 Marks

Solution	On analysis of the plots we can say that there is no such impact of mba_p score and placement of students. This is because the pattern for the number for students scoring a certain marks in mba_p test is same for "Placed" as well as "Not Placed" students, with majority lying in the range of 57 to 63 marks.
	 
	localhost:8888/notebooks/Untitled1.ipynb?kernel_name=python3
	jupyter Untitled1 Last Checkpoint: 43 minutes ago (autosaved)
	File Edit View Insert Cell Kernel Widgets Help
	In [49]: `from matplotlib import pyplot as plot`
	In [61]: `mean_mbap_unplaced, mean_mbap_placed = df.groupby('status')['mba_p']`
	`mean_mbap_placed`
	`mean_mbap_unplaced`
	Out[61]: ('Not Placed',
	3 59.43
	5 51.58
	6 53.29
	9 52.21
	12 65.04
	...
	198 71.96
	201 58.44
	206 53.39
	208 62.92
	214 60.22
	Name: mba_p, Length: 67, dtype: float64)
	In [66]: `placed_plot = plot.hist(mean_mbap_placed[1:])`
	`unplaced_plot = plot.hist(mean_mbap_unplaced[1:])`
	`# placed_plot.show()`
	`# unplaced_plot.show()`

Question 5	You want to understand the impact of academic …..
Marks	8 Marks
5.1	
5.2	

5.3



5.4

Steps Taken: -

1) We calculated the 75th percentile value for all the required values by using .quantile(0.75)
2) Then we filtered all whole dataframe based on the 75th percentile values of the test scores. If a student has score marks equal to or above the 75th percentile marks for that test, he/she is considered as student of High Acumen.
3) We calculated the placement percentage for the students with high acumen.

Since, among the high acumen students, 80 % students got placed we can say that the hypothesis is somewhat true i.e. the high acumen students mostly leads to better placements.