The report for 4140 Assignment 3

1. Application and data sets
1)Data source:
   https://relational.fit.cvut.cz/dataset/CCS
This dataset contains card transactions made by customers when they pay for gas products in the gas stations.
The business scenario:
   The Czech Debit Card Company is an international chain company. Czech debit cards are used to pay for petrol pumps. The company wants to promote the usage of company's debit card and they has cooperated with the gas station for several years. They think that people from different countries have their own consumption habits when they pay for gasoline and it is essential for them to analyze customers' consumption records. Therefore, they have collected all the consumption records of Czech debit cards from cooperated gas stations. But they do not know how to extract useful information from this data. So, they asked us to help them summarize useful information from these data.
1.   Need to know the yearly, monthly and daily transaction for each gas station.
2.   Need to compare transactions in terms of different currency, different chain and segment for gas station, and different segment of customers.
3.   Need to compare the transaction of variable time of the day according to different segments of customers.
4.   Need to know the description for each product, which product has higher demand for each gas station, and the price for each kind of gas product and priced paid by each customer.
   (ER diagram)
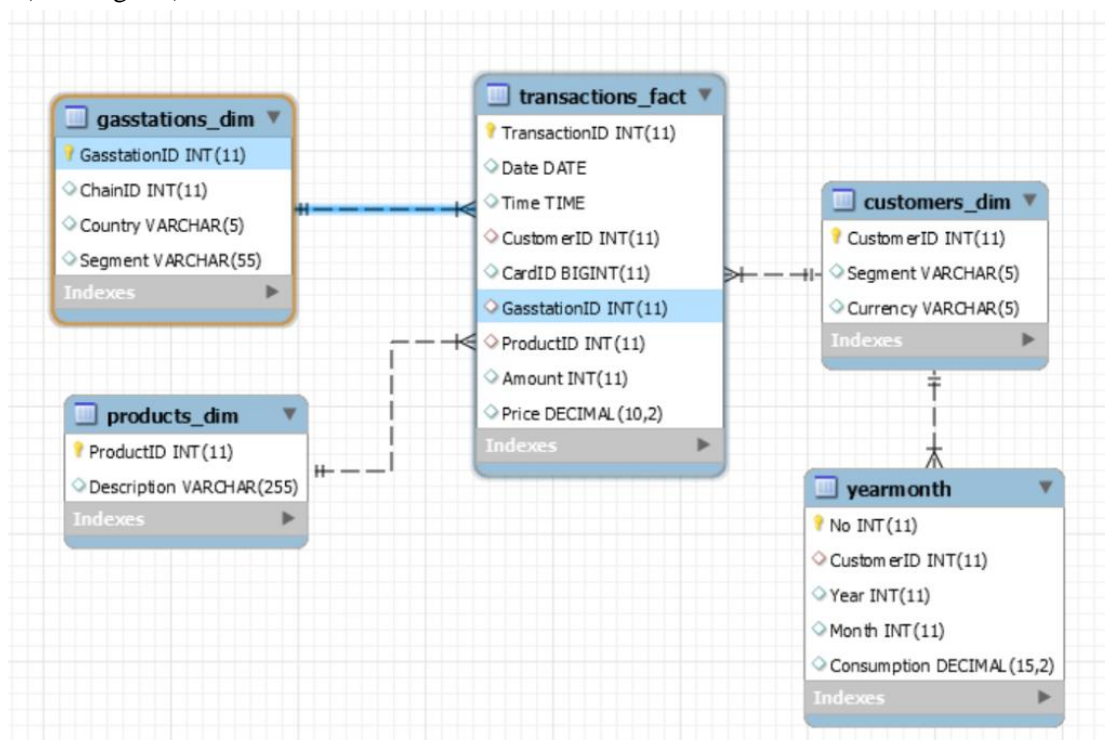
2.Business Queries & Results

1) This query is to list total sales of each gas station which pay by Czech debit card and compare them with their chain and country

select GasstationID,ChainID, sum(Price) as Total, country
from transactions_fact left join gasstations_dim using(GasstationID)
Group by GasstationID
having sum(Price)>5000
order by Total desc, ChainID asc,country asc;

GasstationID	ChainID	Total	country
115	2	10086.86	CZE
434	6	8729.49	CZE
325	2	8715.98	CZE
311	2	6533.26	CZE
339	33	6388.41	CZE
48	4	6338.81	CZE
374	23	5194.73	CZE
312	1	5144.45	CZE
362	1	5031.01	CZE

2) This query is to compare products sales of gas stations in all chain who belong to 'Other' segment in August.

SQL query:
select Segment, Chainid, ProductID, sum(Price) as TotalIncome, sum(Amount) as SaleQuantity
from (select * from transactions_fact inner join gasstations_dim using (gasstationid)
where segment='Other')as t1
group by chainID,productID
having sum(Price) >500
order by ChainID,ProductID;

Segment	ChainID	ProductID	TotalIncome	SaleQuantity
Other	3	2	10862.58	470
Other	3	5	1872.00	80
Other	10	2	8582.22	376
Other	10	5	542.86	23
Other	22	2	2890.14	124
Other	23	2	3303.81	147
Other	23	9	1493.81	61
Other	26	2	1162.48	52
Other	34	2	4105.10	187
Other	36	2	11536.15	506
Other	36	5	779.07	34
Other	36	9	996.08	38
Other	243	2	1272.58	55
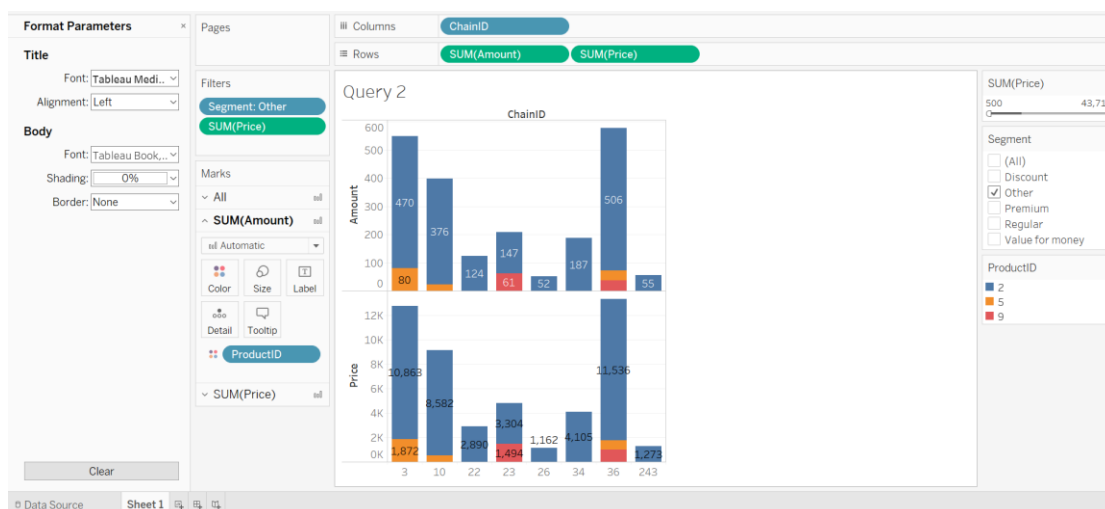
This query is to find in the segment "Other", which products have sales over 500 and which chain they belong to. The segment and the chain ID are attributes of the table gasstations_dim. Thus, the table is a join combination of transactions_fact and gasstations_dim. In the graph, the columns are chainID while the rows are amount and price.

The first chart represents that in Chain 3, the sale of product 2 is 10863 and the amount is 470. The orange part is product 5 which is in the same chain. Its sale is 1872 with amount of 80.

The second chart is chain 10. The product 2's sale is 8582 with amount of 376 while product 5 sold 23 and the income is 543.

The third chart shows that product 2 is the only product in Chain 22 sold over 500 (sold 124 and the income is 2890)

The fourth chart shows in chain 23, product 2 sold 147 and made income 3304 and product 9 sold 61 and made income 1494.

The fifth, sixth, and eighth charts show in chain 26, 34, and 243, product 2 made sale of 52,187, 55 and the income was 1162, 4105, and 1273.

The seventh chart shows in chain 34, the product 2 has the highest sales. It sold 506 and the income was 11536. Product 5 and 9 have close sales. The amount was 34 and 38 while the income is 779 and 996.

Through this analysis, it is clear that product 2 has occupied the market in the "Other" segment. In addition, product 5 and 9 has potential market.

3) This query is to find product 2's peak consuming hour of a day and is grouped by customers' segment. (only include customers using CZK currency)

SQL query:

select productid,segment,sum(amount) as TotalSale,currency,left(time,2) as 'Hour' from transactions_fact inner join customers_dim using (customerID) where Currency like 'CZK' and ProductID=2
group by Hour,segment
order by Hour,segment desc;

productid	segment	TotalSale	currency	Hour
2	SME	42	CZK	00
2	SME	79	CZK	02
2	SME	88	CZK	03
2	KAM	53	CZK	03
2	SME	236	CZK	04
2	SME	219	CZK	05
2	LAM	22	CZK	05
2	SME	265	CZK	06
2	KAM	27	CZK	06
2	SME	273	CZK	07
2	LAM	46	CZK	07
2	KAM	44	CZK	07
2	SME	565	CZK	08
2	LAM	69	CZK	08
2	KAM	100	CZK	08
2	SME	631	CZK	09
2	LAM	155	CZK	09
2	KAM	51	CZK	09
2	SME	342	CZK	10
2	LAM	138	CZK	10
2	KAM	76	CZK	10
2	SME	324	CZK	11
2	LAM	213	CZK	11
2	KAM	52	CZK	11
2	SME	498	CZK	12
2	LAM	585	CZK	12
2	KAM	84	CZK	12
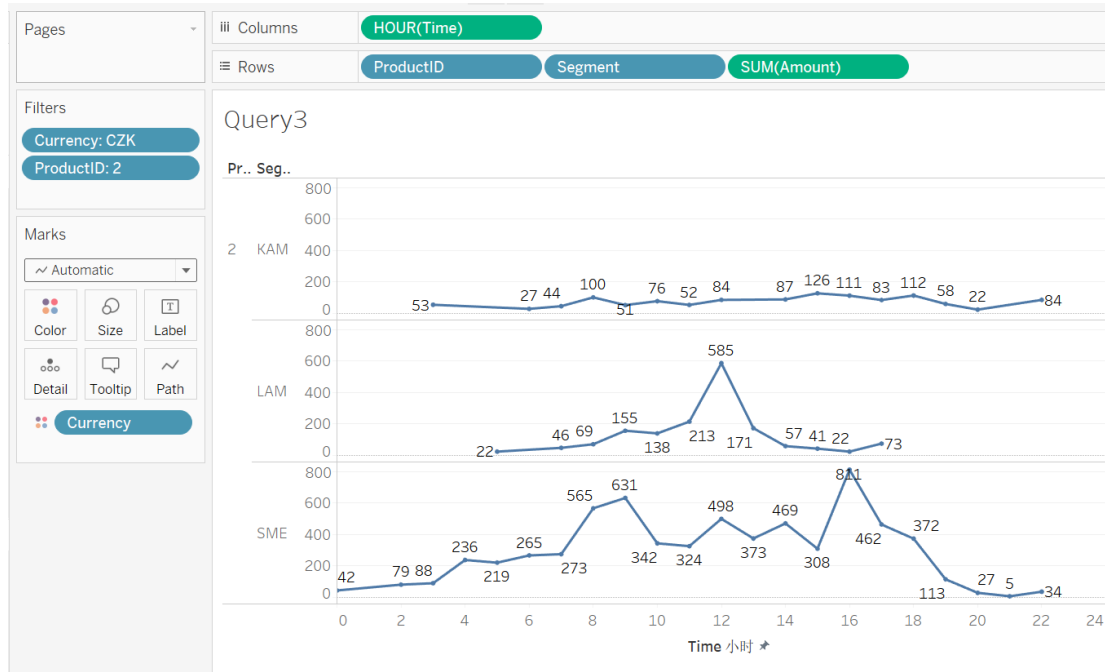
Each color represents a segment of the customer. Through the analysis, we can see that 9.am,12 pm, and 4 pm are peak selling hours for segment SME, 8 am, 3 pm, 6 pm, and 10 pm for segment KAM, 12 pm for segment LAM.

4) This query is to find the total consumption of each consumer, and to compare them in terms of Currency and segment
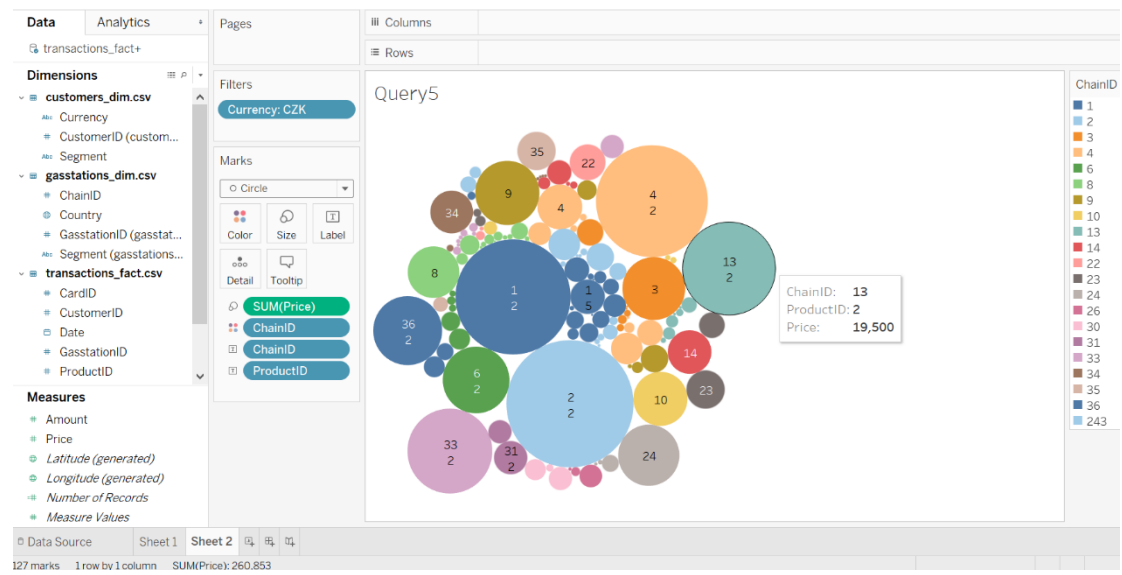
```
select Currency,Segment,CustomerID,    sum(Consumption) as Sum_Consump
  from yearmonth_dim left join customers_dim using (customerID)
  Group by Currency asc,Segment asc,CustomerID desc
  having sum(Consumption)>50000
  order by Currency asc,sum(Consumption) desc;
```

Currency	Segment	CustomerID	Sum_Consump
CZK	LAM	335	554257.63
CZK	SME	331	213184.43
CZK	SME	314	182491.23
CZK	SME	324	178808.10
CZK	SME	305	71463.04
CZK	LAM	323	63170.04
EUR	LAM	51	322303.73
EUR	KAM	105	189352.01
EUR	SME	38	120050.53
EUR	SME	289	115206.04
EUR	LAM	272	112714.48
EUR	LAM	96	85358.76
EUR	LAM	122	75043.24
EUR	KAM	285	73419.53
EUR	KAM	154	69064.44
EUR	LAM	235	67275.49
EUR	LAM	224	59430.16
EUR	LAM	172	51473.47

5）This query finds popular product in different chains. Through such query, we can learn which products attract more customers and their comparsion among different chains.

select sum(Price) as TotalPrice,chainID,ProductID from(

select * from transactions_fact inner join (select customerID,currency from customers_dim) as t using(customerID) inner join gasstations_dim using (gasstationID) where Currency='CZK') as t1

group by chainID,ProductID

order by sum(Price) desc, ChainID asc,ProductID asc;

TotalPrice	ChainID	ProductID
36411.27	2	2
29837.16	1	2
28205.96	4	2
19500.49	13	2
16214.45	33	2
10707.99	36	2
10021.60	6	2
9278.60	9	2
8900.83	3	2
8428.74	24	2
6589.06	10	2
5993.64	8	2
4493.41	4	9
4280.90	14	2
4105.10	34	2
3308.44	35	2
3303.81	23	2
2890.14	22	2
2676.89	1	5
2502.69	31	2
2242.25	4	5
2184.49	2	8
1718.04	2	5
1705.38	9	20
1583.61	3	5
1493.81	23	9
1435.83	4	31

3. Summary