Final Report on Big Data Analysis Course – Case 3

1. Case Description

I have selected Case 3 from the cases suggested for the final report. Case 3 consists of information about clothes sales. Data is stored in 4 sheets in XLS file.

Sheet 1 – Orders – consists of 14 columns and 2172 rows, main financial information regarding sales is stored on this sheet.

- OrderID – order identification,
- OrderDate,
- Month,
- Year,
- CustomerID,
- EmployeeID,
- ShipperID,
- ProductID,
- Sales,
- Costs,
- Profit,
- Quantity,
- Discount,
- Freight

Sheet 2 – Categories – consists of 3 columns и 8 rows. This sheet contains information about categories of the goods that are being sold.

- CategoryID,
- Category,
- Description.

Sheet 3 – Products – consists of 4 columns and 77 rows, the sheet contains information about the products that are being sold.

- ProductID,
- CategoryID,
- Product,
- SupplierID.

Sheet 4 – Suppliers – consists of 5 columns and 29 rows, this sheet contains information about suppliers.

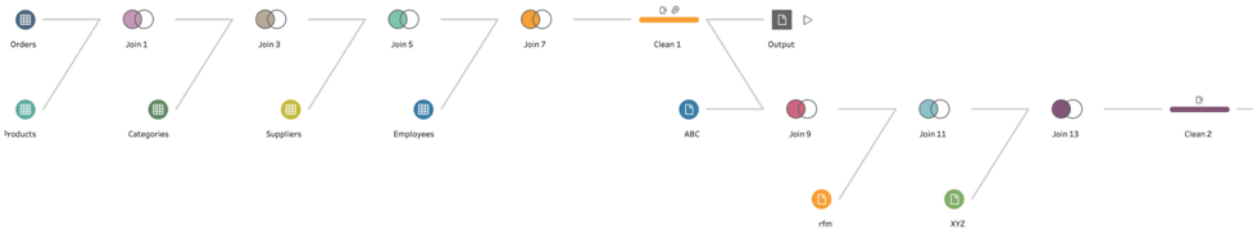
- ID,
- SupplierID,
- Supplier,
- SupplierContact,
- SupplierCountry.

Sheet 5 – Employees – consists of 8 columns and 47 rows, the sheet contains information about employees.

- EmployeeID,
- Extension,
- EmployeeName,
- HireDate,
- Office,
- ReportsTo,
- Title,
- YearSalary.

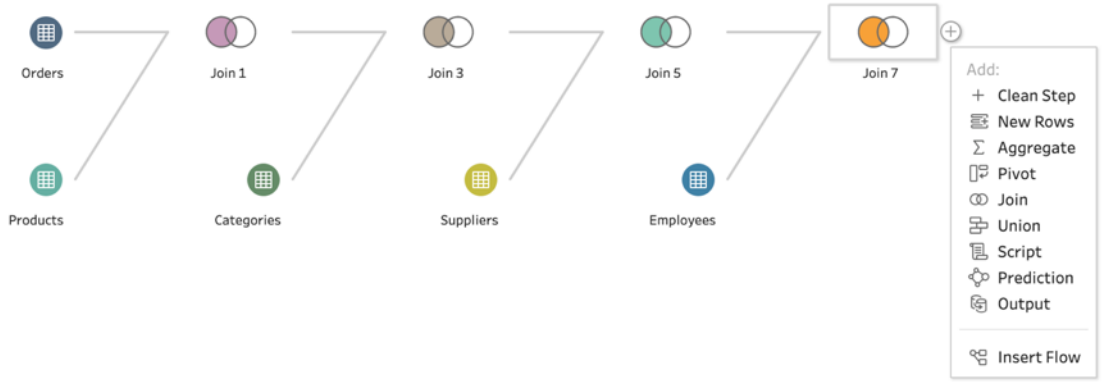
2. Loading the data, consolidating and EDA using Tableau Prep Builder

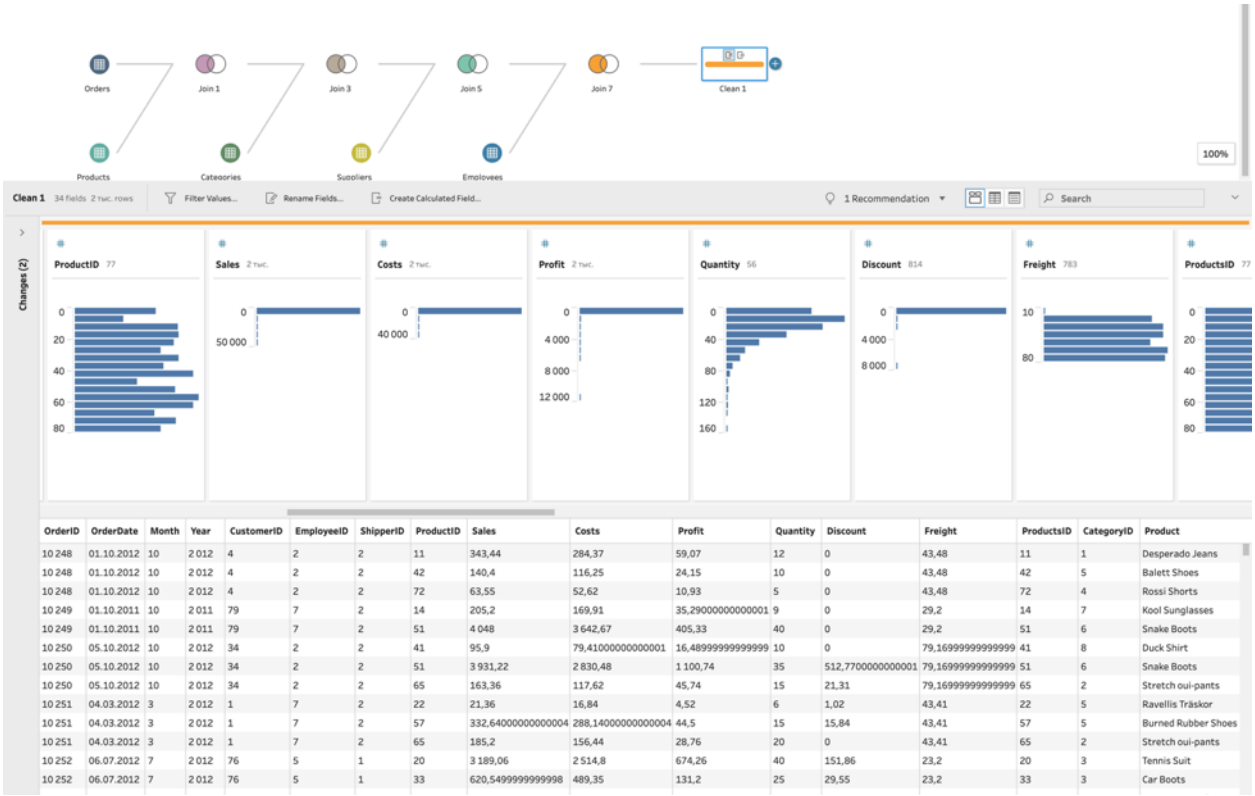
Tables concatenation was done using Tableau Prep Builder. Data scheme is shown in the picture below.



Data consolidation and cleaning

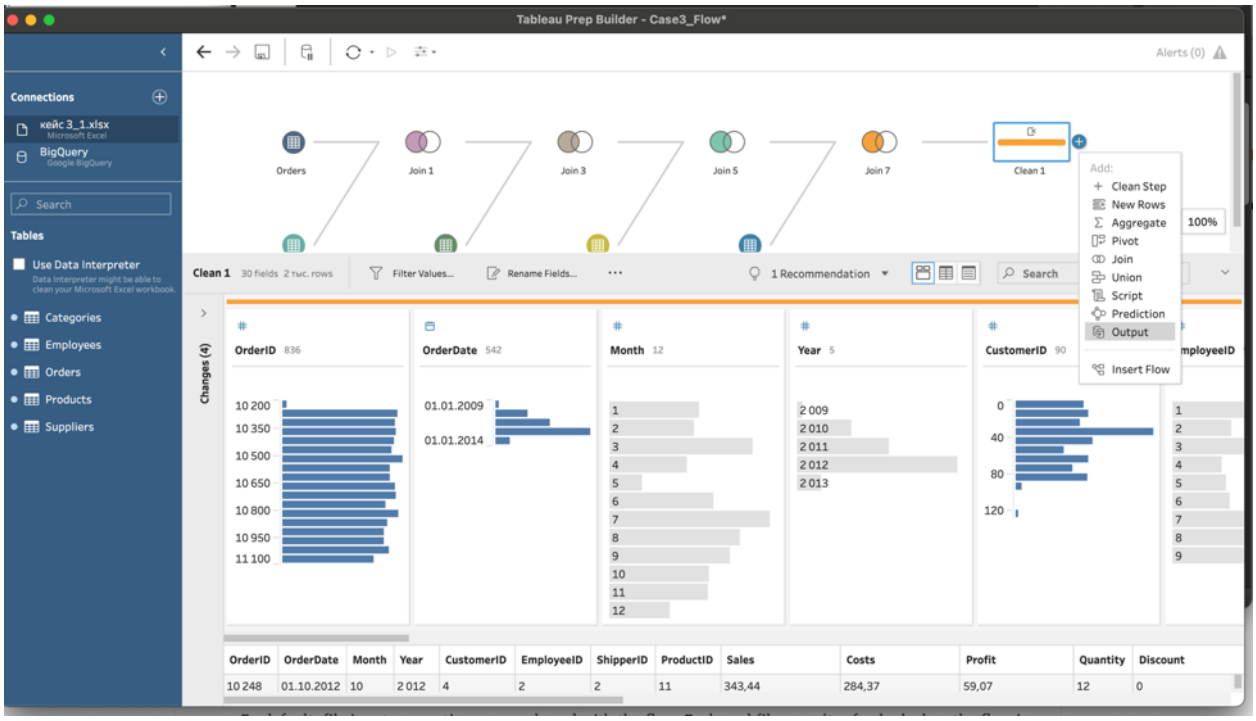
We check the data and remove repeated columns using Clean.

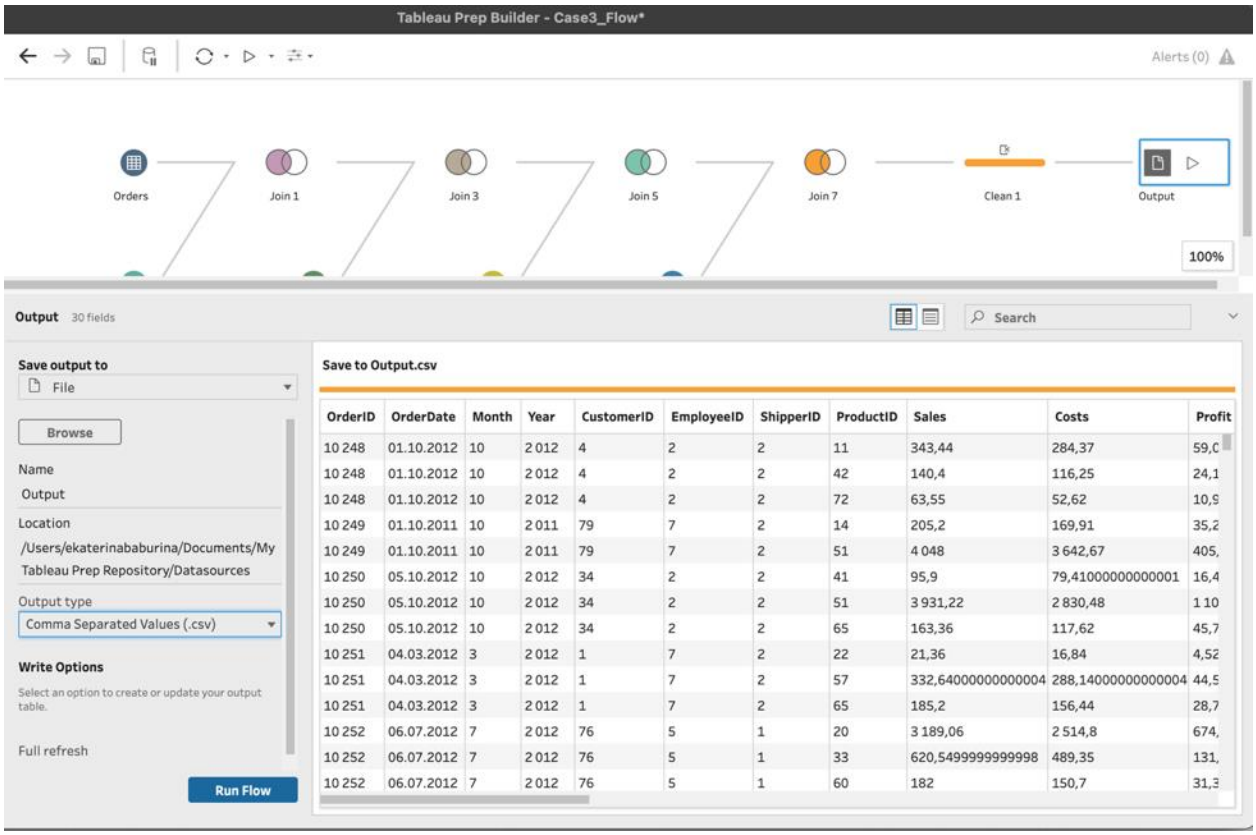




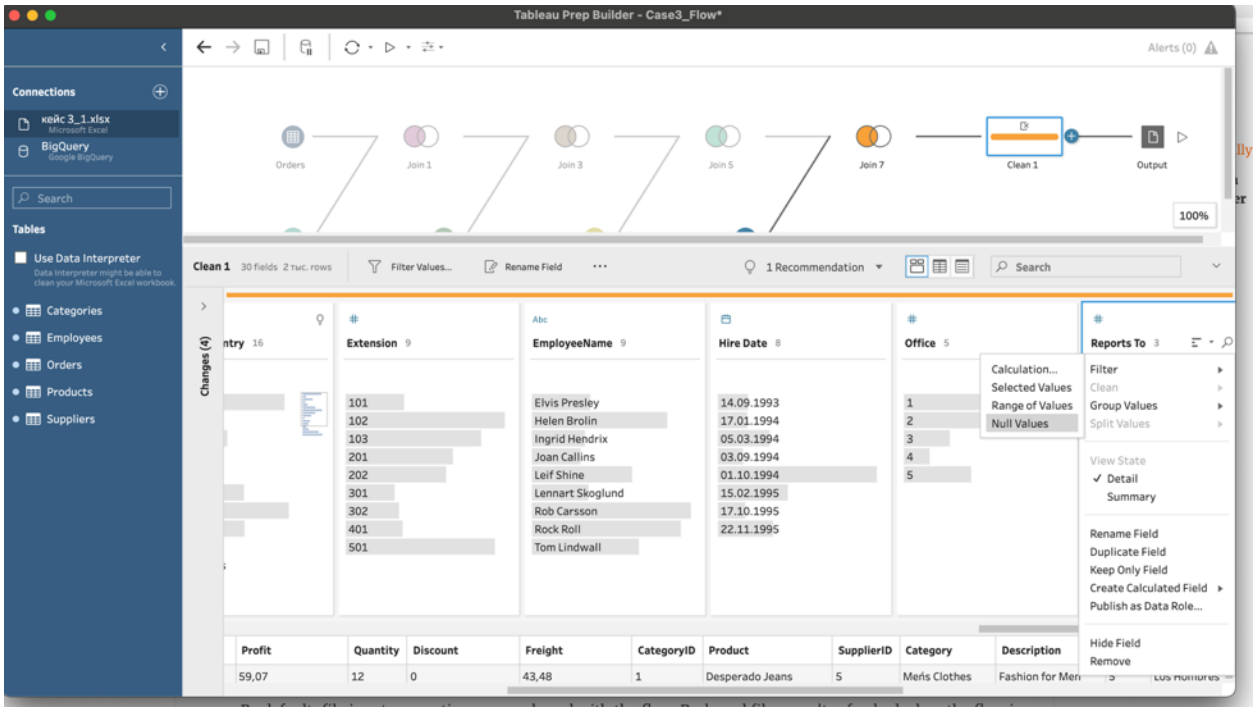
OrderID	OrderDate	Month	Year	CustomerID	EmployeeID	ShipperID	ProductID	Sales	Costs	Profit	Quantity	Discount	Freight	ProductsID	CategoryID	Product
10 248	01.10.2012	10	2 012	4	2	2	11	343,44	284,37	59,07	12	0	43,48	11	1	Desperado Jeans
10 248	01.10.2012	10	2 012	4	2	2	42	140,4	116,25	24,15	10	0	43,48	42	5	Balett Shoes
10 248	01.10.2012	10	2 012	4	2	2	72	63,55	52,62	10,93	5	0	43,48	72	4	Rossi Shorts
10 249	01.10.2011	10	2 011	79	7	2	14	205,2	169,91	35,29000000000001	9	0	29,2	14	7	Kool Sunglasses
10 249	01.10.2011	10	2 011	79	7	2	51	4 048	3 642,67	405,33	40	0	29,2	51	6	Snake Boots
10 250	05.10.2012	10	2 012	34	2	2	41	95,9	79,41000000000001	16,48999999999999	10	0	79,16999999999999	41	8	Duck Shirt
10 250	05.10.2012	10	2 012	34	2	2	51	3 931,22	2 830,48	1 100,74	35	512,7700000000001	79,16999999999999	51	6	Snake Boots
10 250	05.10.2012	10	2 012	34	2	2	65	163,36	117,62	45,74	15	21,31	79,16999999999999	65	2	Stretch oui-pants
10 251	04.03.2012	3	2 012	1	7	2	22	21,36	16,84	4,52	6	1,02	43,41	22	5	Ravellis Träskor
10 251	04.03.2012	3	2 012	1	7	2	57	332,64000000000004	288,14000000000004	44,5	15	15,84	43,41	57	5	Burned Rubber Shoes
10 251	04.03.2012	3	2 012	1	7	2	65	185,2	156,44	28,76	20	0	43,41	65	2	Stretch oui-pants
10 252	06.07.2012	7	2 012	76	5	1	20	3 189,06	2 514,8	674,26	40	151,86	23,2	20	3	Tennis Suit
10 252	06.07.2012	7	2 012	76	5	1	33	620,5499999999998	489,35	131,2	25	29,55	23,2	33	3	Car Boots

We add Output and export the file to csv.

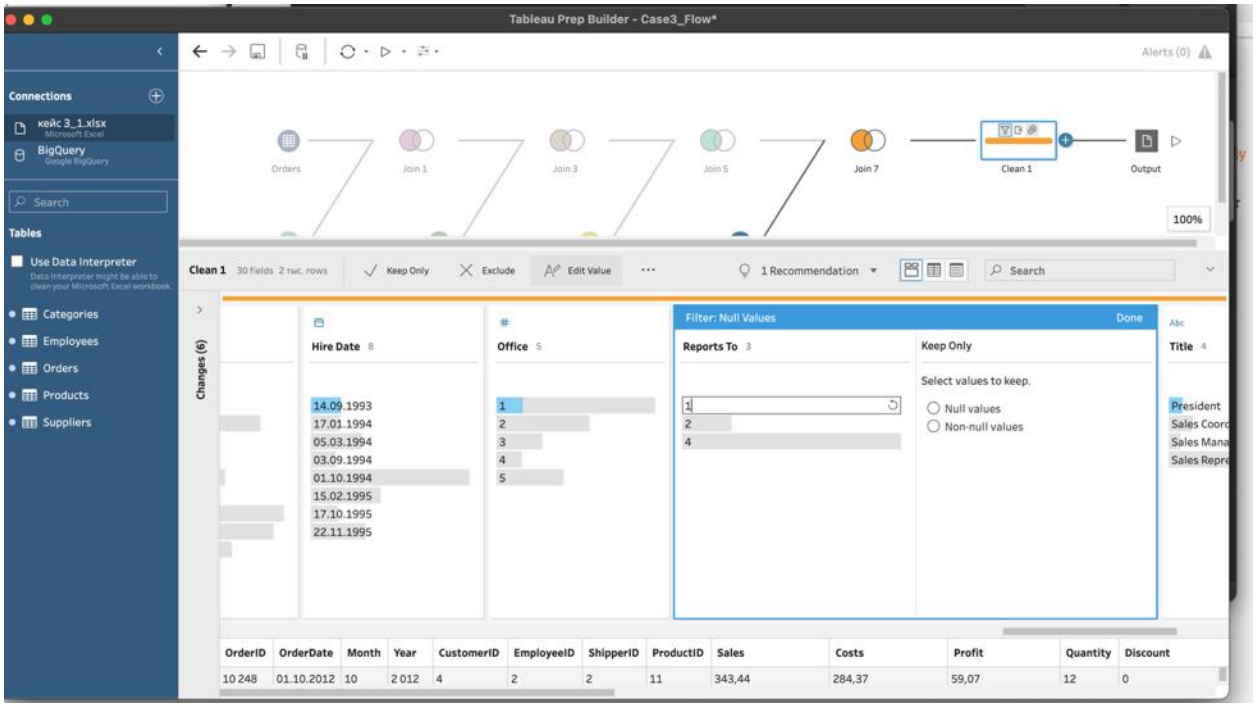


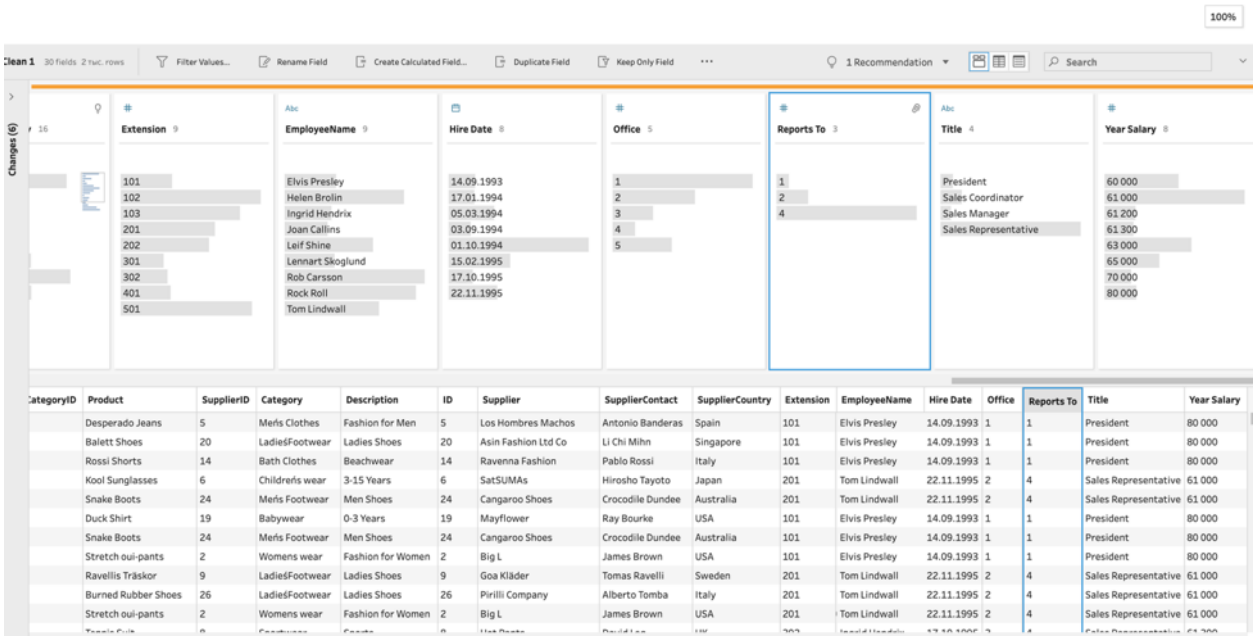


EDA is needed, we need to check for missing values and check if the feature's format is correct.

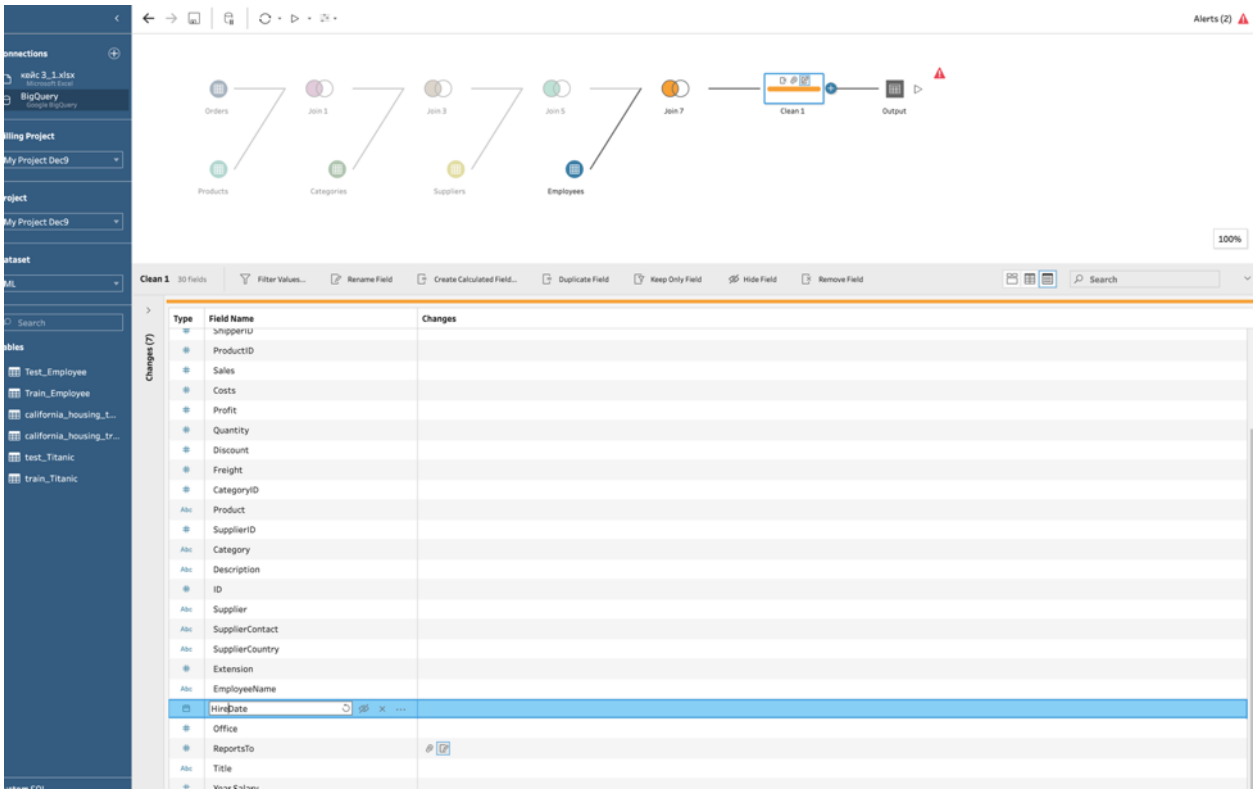


We can see that there are **Null** values in "**Reports To"** field, so we will replace null values with "**1**" to be able to build the model.





Exporting clean table to **Google Big Query**.



3. Data enrichment with ABC, XYZ, RFM analysis in Google BigQuery

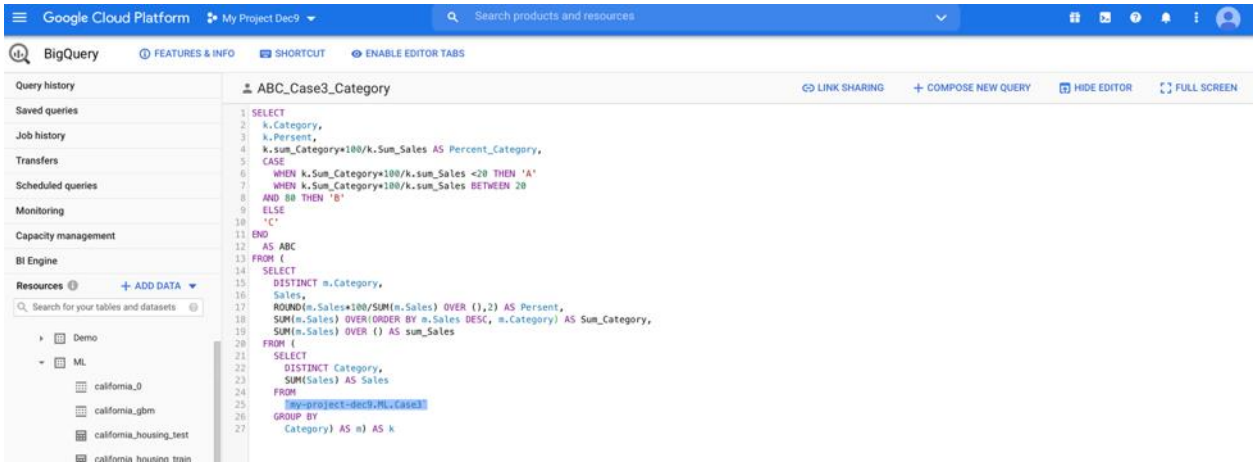
3.1. Adding ABC analysis to the dataset

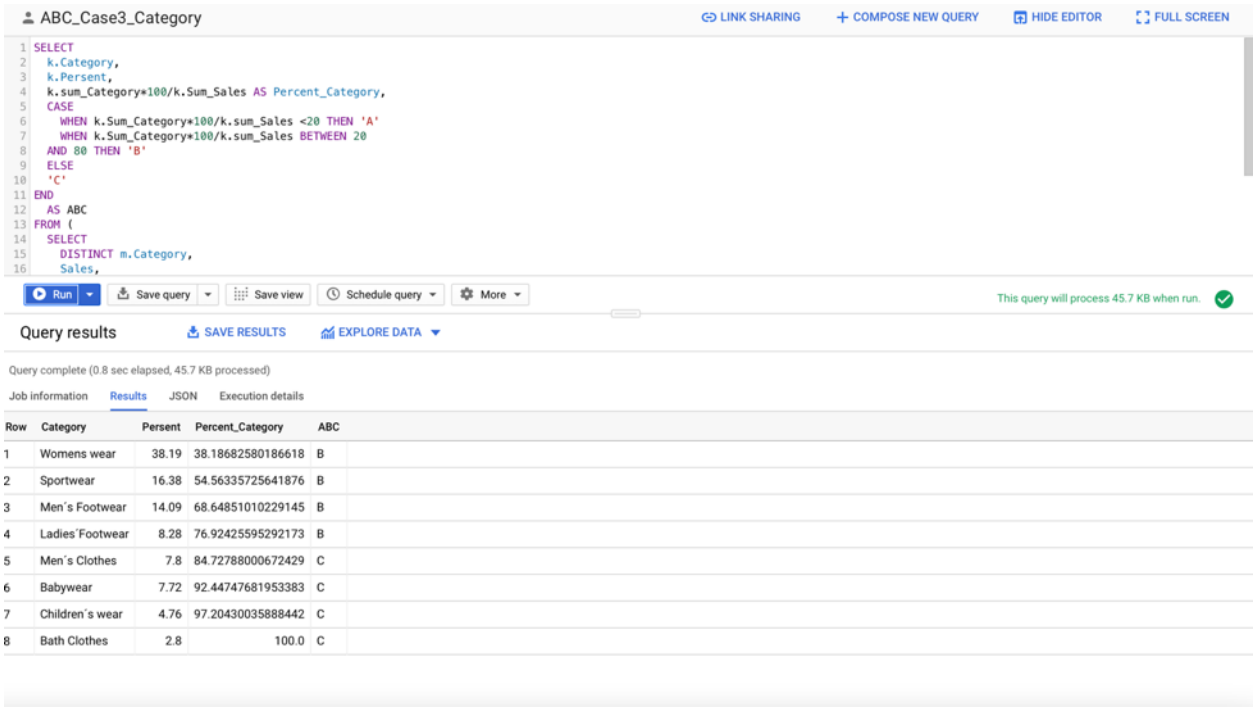
ABC analysis is an inventory categorization technique.
As the name implies, ABC analysis sorts inventory into three main buckets:

- **A items**: This is your inventory with the highest annual consumption value. It should be your highest priority and rarely, if ever, a stockout. 20% of the items accounts for 80% in sales.

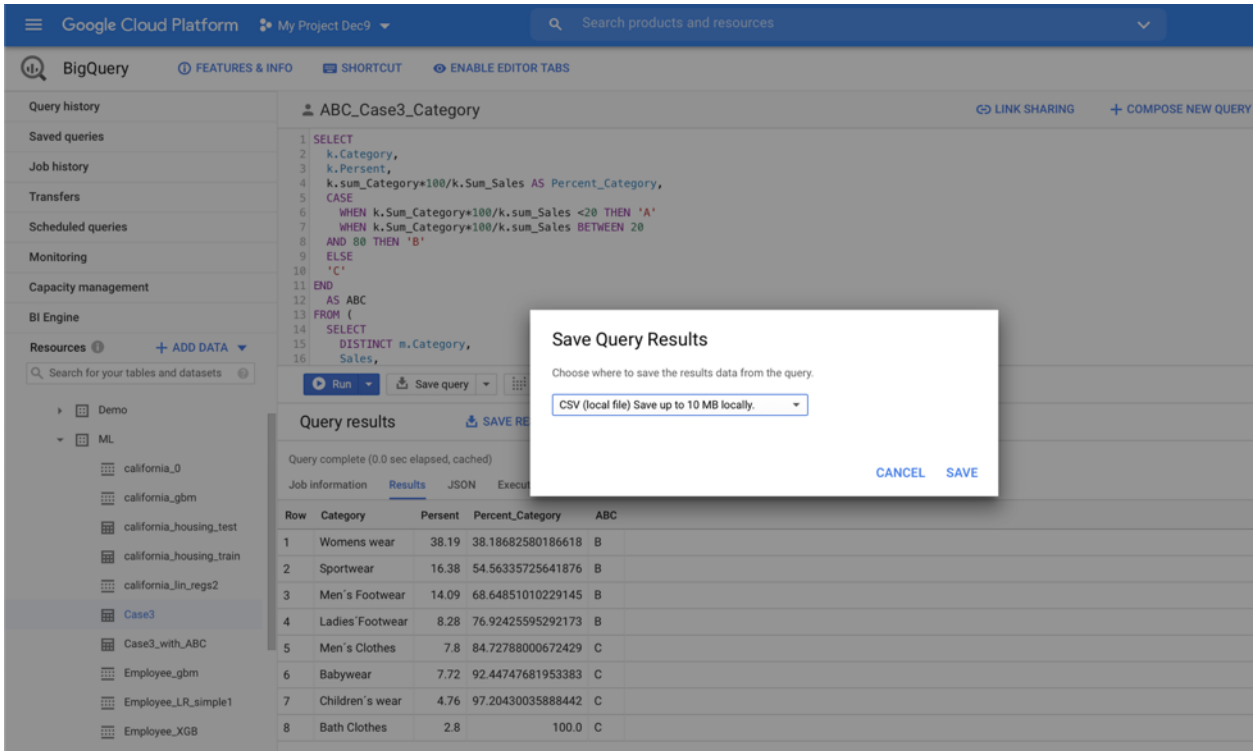
- **B items:** Inventory that sells regularly but not nearly as much as A items. Often inventory that costs more to hold than A items. 30% of the items accounts for 15% in sales.
- **C items:** This is the rest of your inventory that doesn't sell much, has the lowest inventory value, and makes up the bulk of your inventory cost. 50% of the items accounts for 5% in sales.

Inventory categorization is essential with physical products because it protects the profit margins and prevents write-offs and losses for spoiled inventory. It is also the first step in reducing obsolete inventory, supply chain optimization, increasing prices, and forecasting demand.





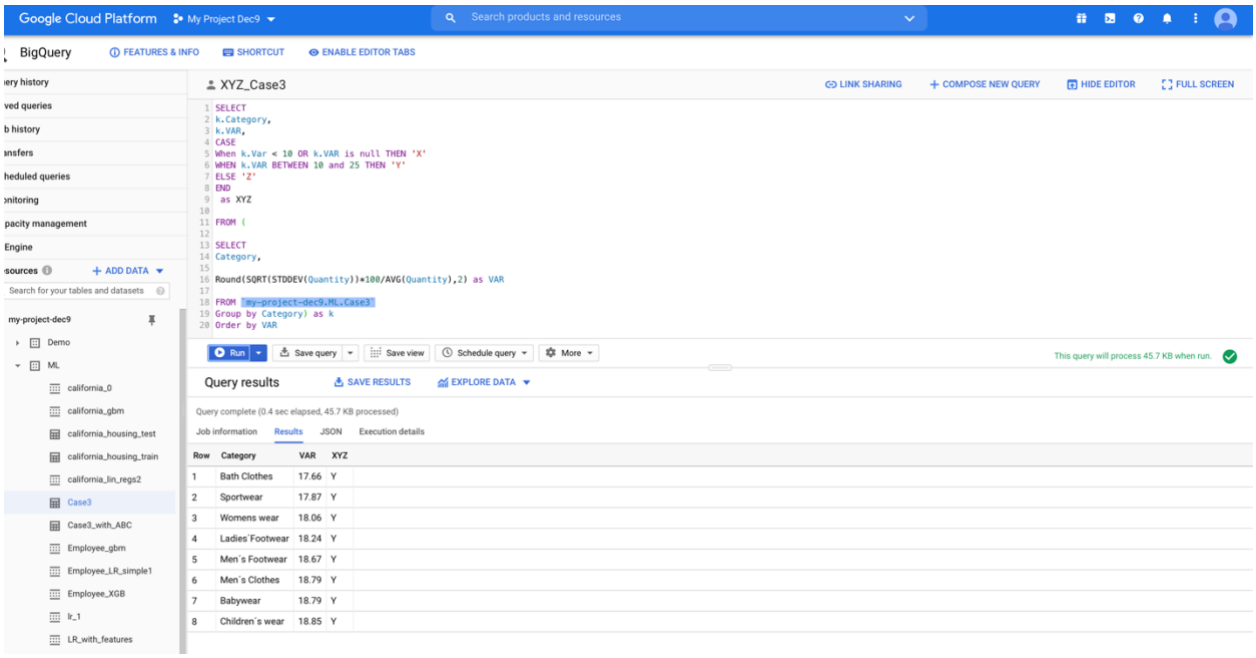
We can see that in the dataset provided for this test case only B and C categories are present. We save the results of ABC analysis to file for further usage.



3.2. Adding XYZ analysis to the dataset

The **XYZ analysis** is a way to classify inventory items according to variability of their demand.

- **X – Very little variation**: X items are characterized by steady turnover over time. Future demand can be reliably forecasted. The coefficient of variation is between 0% and 10%.
- **Y – Some variation**: Although demand for Y items is not steady, variability in demand can be predicted to an extent. This is usually because demand fluctuations are caused by known factors, such as seasonality, product lifecycles, competitor action or economic factors. It's more difficult to forecast demand accurately. The coefficient of variation is between 10% and 25%.
- **Z – The most variation**: Demand for Z items can fluctuate strongly or occur sporadically. There is no trend or predictable causal factors, making reliable demand forecasting impossible. The coefficient of variation is above 25%.



We are saving the results for further use.

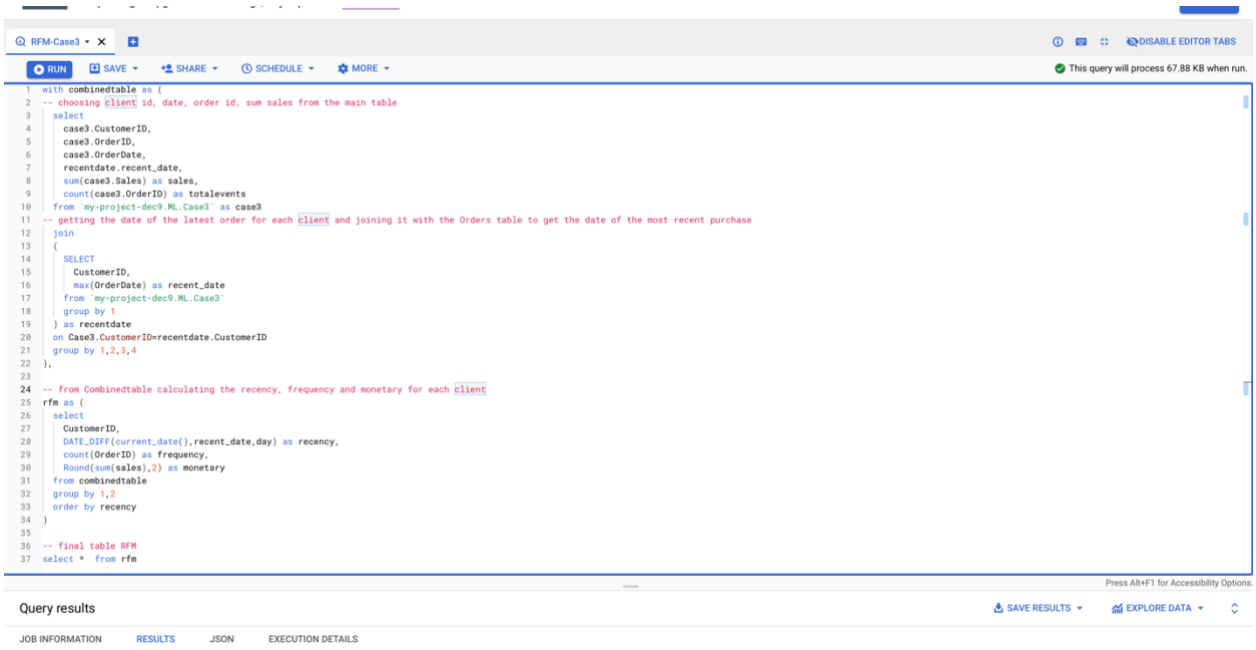
3.3. Adding RFM analysis

RFM analysis is a marketing technique used to quantitatively rank and group customers based on the **recency, frequency and monetary** total of their recent transactions to identify the best customers and perform targeted marketing campaigns. The system assigns each customer numerical scores based on these factors to provide an objective analysis. RFM analysis is based on the marketing adage that "80% of your business comes from 20% of your customers." RFM analysis ranks each customer on the following factors:

- **Recency**. How recent was the customer's last purchase? Customers who recently made a purchase will still have the product on their mind and are more likely to purchase or use the product again. Businesses often measure recency in days. But, depending on the product, they may measure it in years, weeks or even hours.
- **Frequency.** How often did this customer make a purchase in a given period? Customers who purchased once are often are more likely to purchase again. Additionally, first time customers may be good targets for follow-up advertising to convert them into more frequent customers.

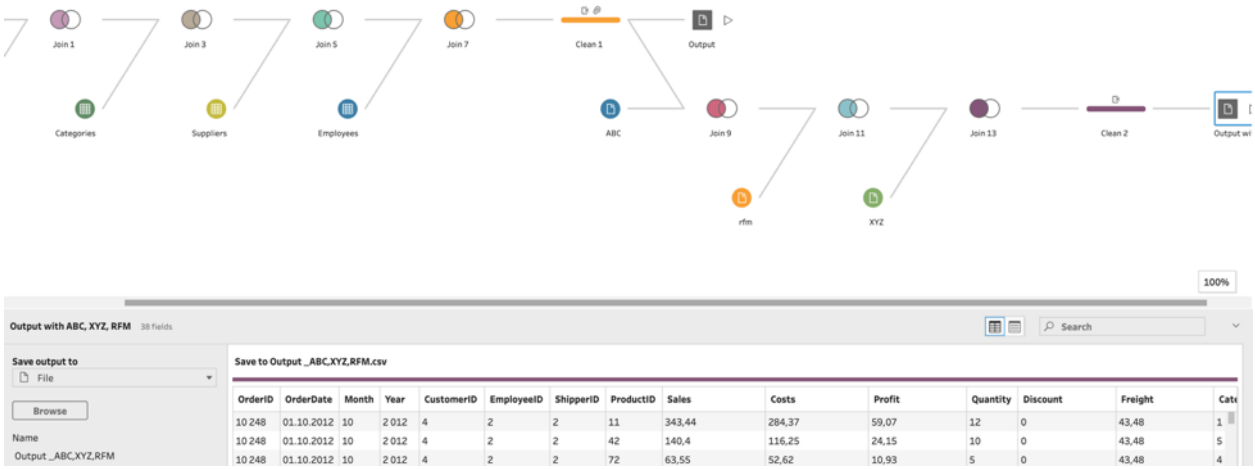
- **Monetary.** How much money did the customer spend in a given period? Customers who spend a lot of money are more likely to spend money in the future and have a high value to a business.

The date of the last order in the test dataset is **29.03.2013**, so we will use this date in **RFM analysis** as the date we calculate the days for **Recency** factor.



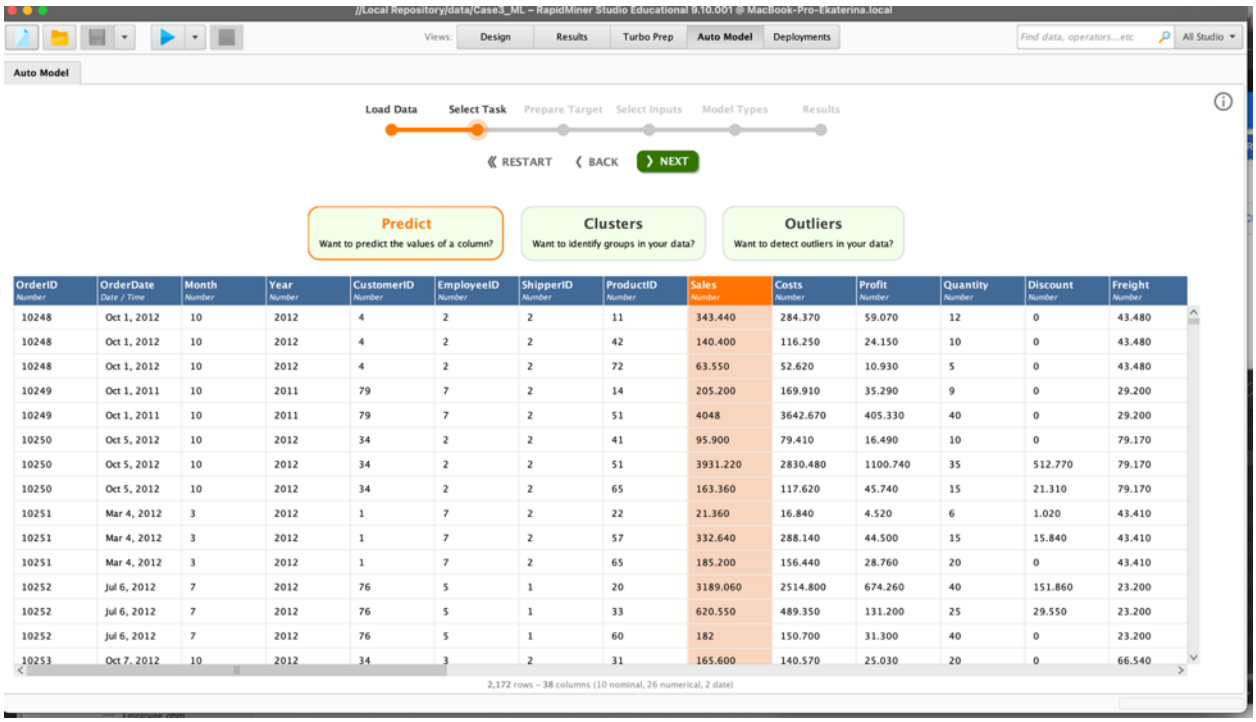
Then the results are saved to the CSV file.

Consolidating the clean file (tables consolidated) and **ABC, XYZ, RFM analysis** in **Tableau Prep Builder.**

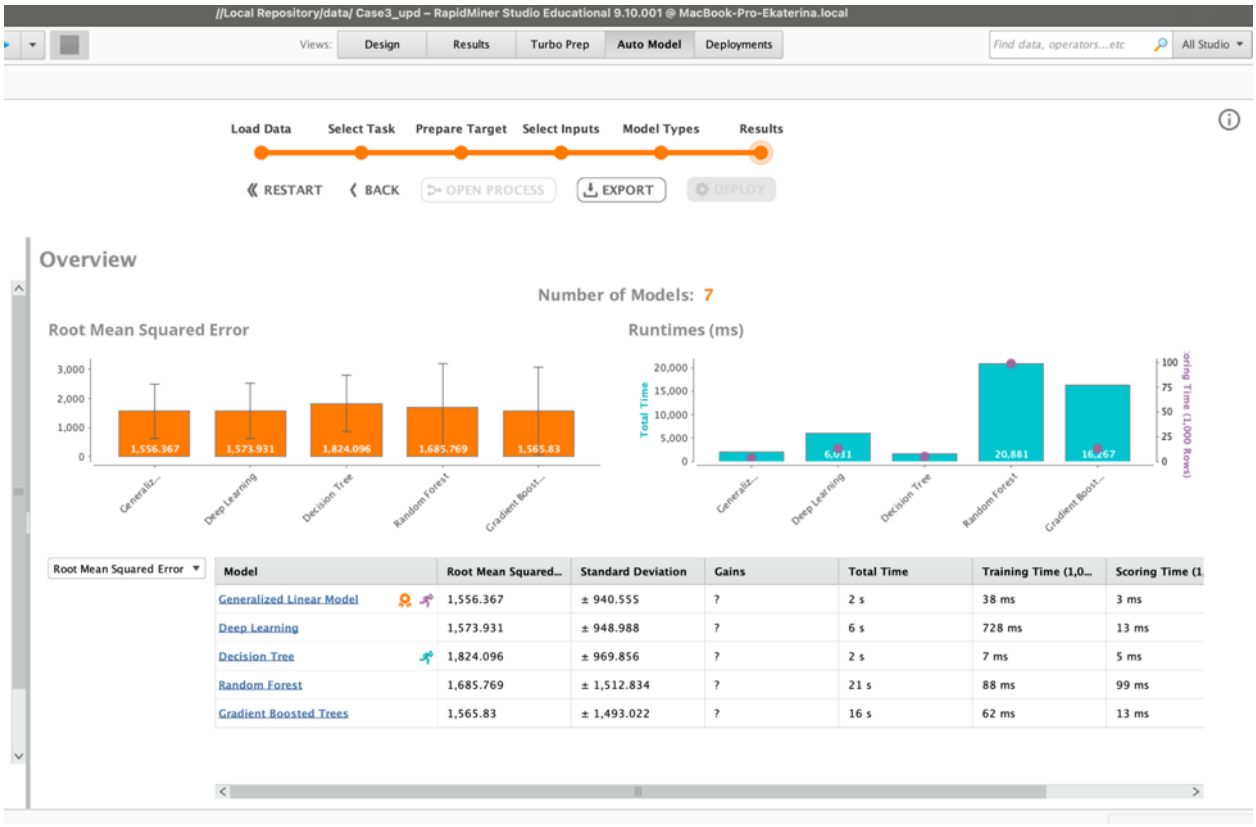


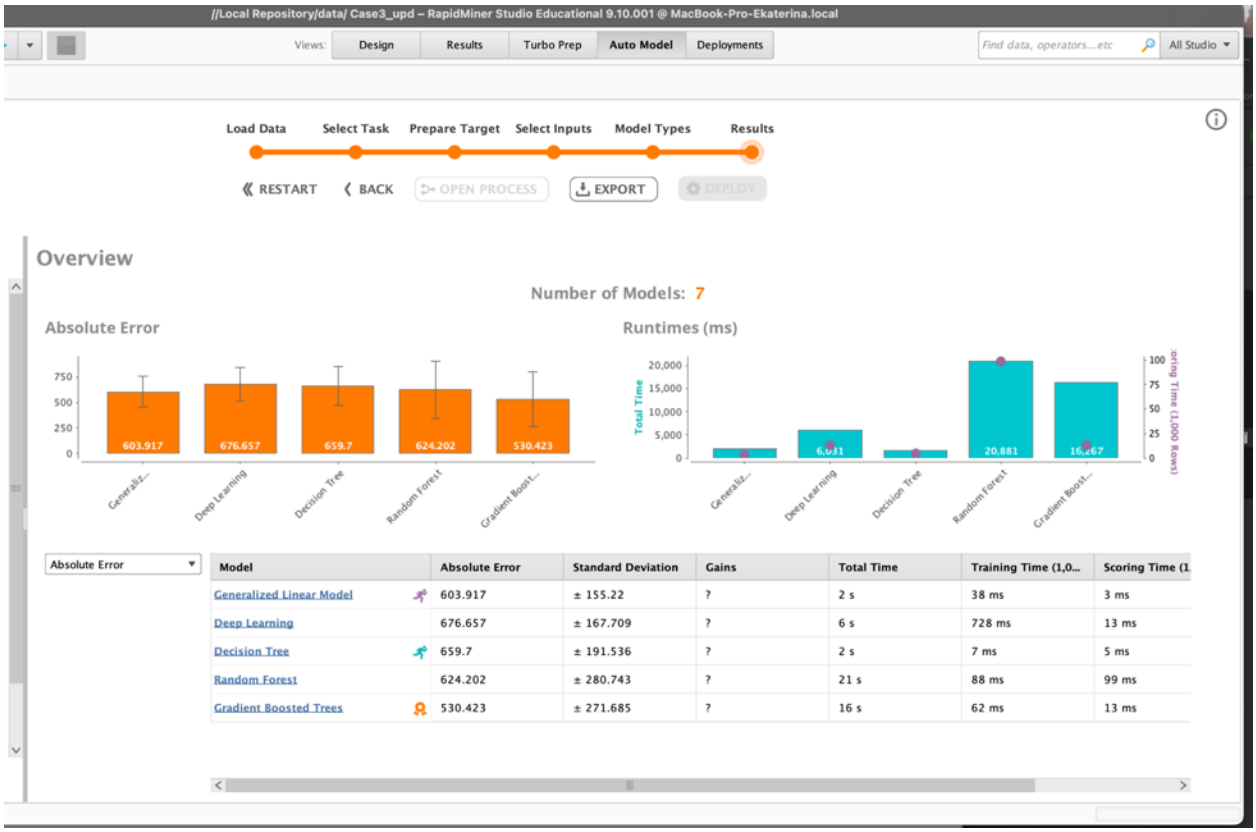
4. Building the model using RapidMiner

Opening the cleaned file in **RapidMiner** and creating the **Machine Learning** model.
We are creating the ML model to predict the **Sales Volumes**.



The next step is to evaluate the models. We will use **RMSE, MAE** criteria to compare the models. And the best model is **Generalized Linear Model.**



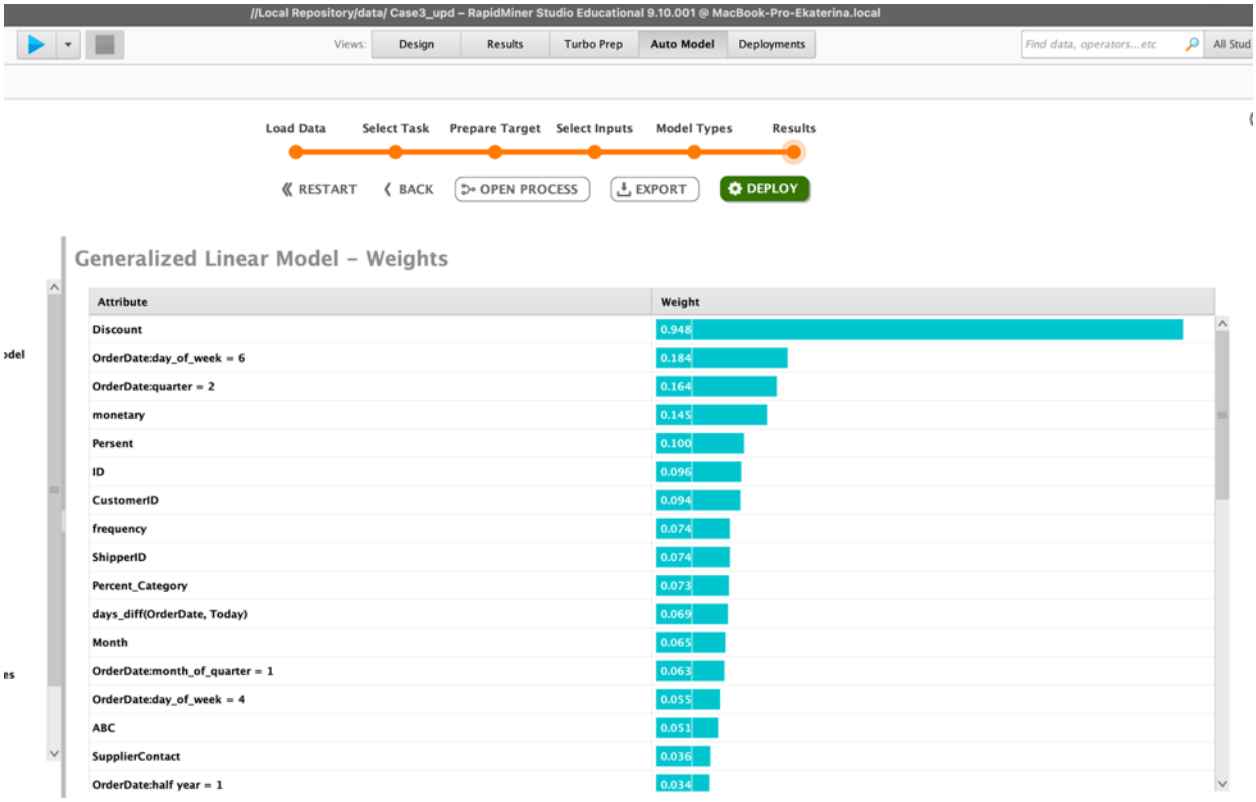


root_mean_squared_error
root_mean_squared_error: 1556.367 +/- 940.555 (micro average: 1769.179 +/- 0.000)

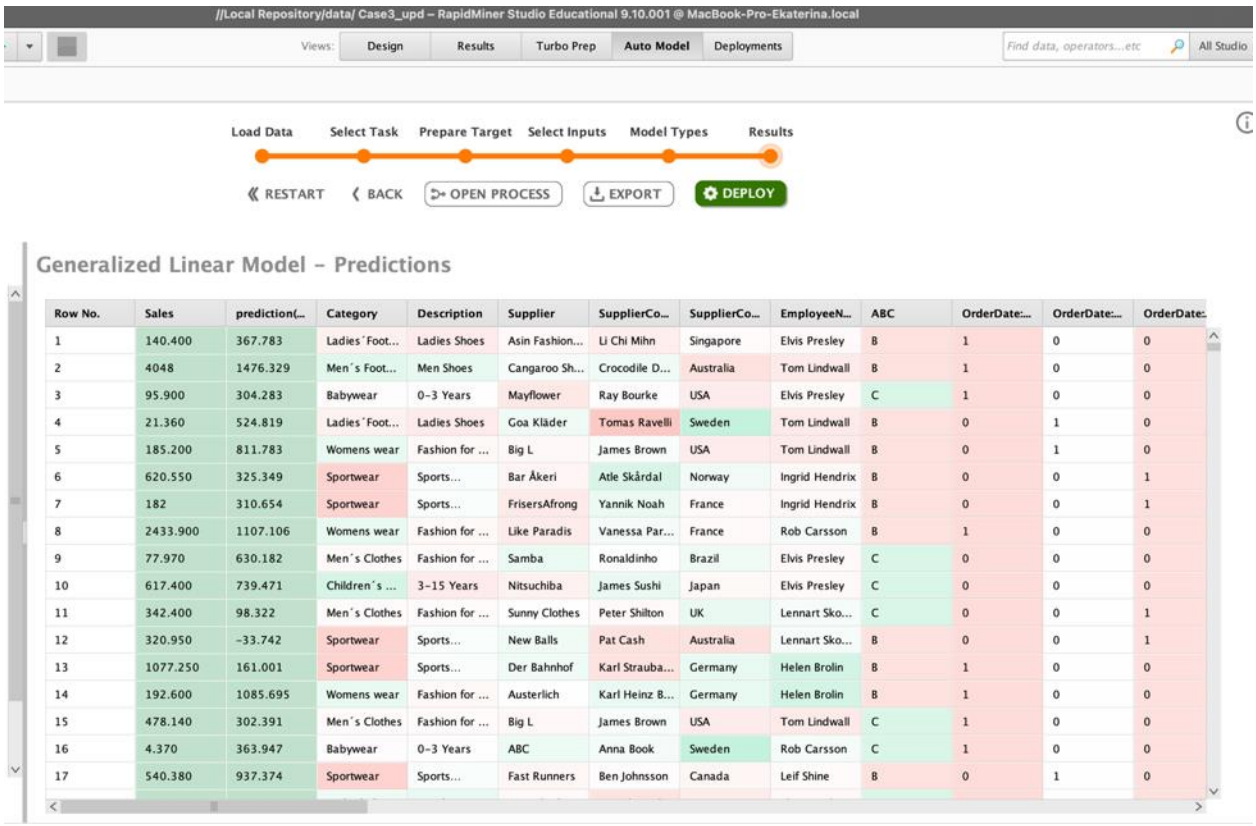
absolute_error
absolute_error: 603.917 +/- 155.220 (micro average: 603.917 +/- 1662.912)

The most important **features** in this model are the following features:

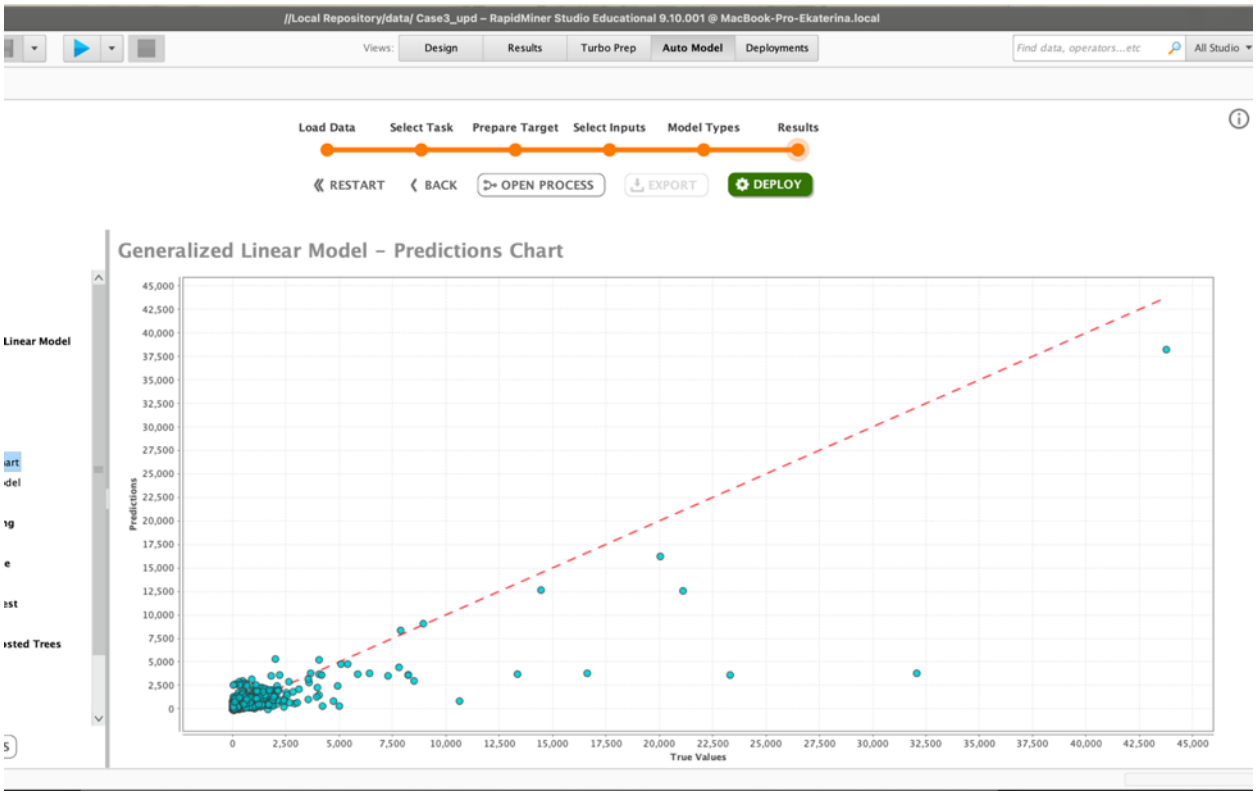
- **Discount,**
- **Day of Order,**
- **Quarter the Order was placed,**
- **Monetary feature of RFM analysis**.



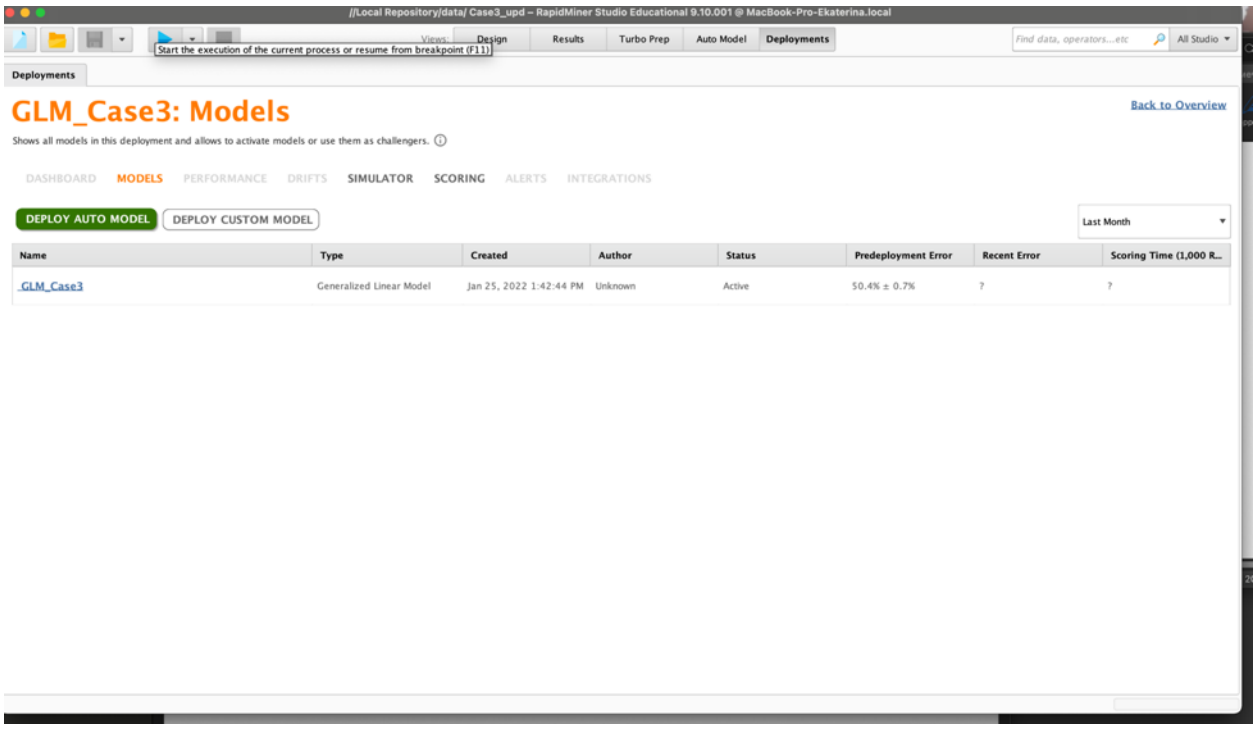
Model’s predictions are shown below:



Row No.	Sales	prediction(...	Category	Description	Supplier	SupplierCo...	SupplierCo...	EmployeeN...	ABC	OrderDate:...	OrderDate:...	OrderDate:
1	140.400	367.783	Ladies´Foot...	Ladies Shoes	Asin Fashion...	Li Chi Mihn	Singapore	Elvis Presley	B	1	0	0
2	4048	1476.329	Men´s Foot...	Men Shoes	Cangaroo Sh...	Crocodile D...	Australia	Tom Lindwall	B	1	0	0
3	95.900	304.283	Babywear	0-3 Years	Mayflower	Ray Bourke	USA	Elvis Presley	C	1	0	0
4	21.360	524.819	Ladies´Foot...	Ladies Shoes	Goa Kläder	Tomas Ravelli	Sweden	Tom Lindwall	B	0	1	0
5	185.200	811.783	Womens wear	Fashion for ...	Big L	James Brown	USA	Tom Lindwall	B	0	1	0
6	620.550	325.349	Sportwear	Sports...	Bar Åkeri	Atle Skårdal	Norway	Ingrid Hendrix	B	0	0	1
7	182	310.654	Sportwear	Sports...	FrisersAfrong	Yannik Noah	France	Ingrid Hendrix	B	0	0	1
8	2433.900	1107.106	Womens wear	Fashion for ...	Like Paradis	Vanessa Par...	France	Rob Carsson	B	1	0	0
9	77.970	630.182	Men´s Clothes	Fashion for ...	Samba	Ronaldinho	Brazil	Elvis Presley	C	0	0	0
10	617.400	739.471	Children´s ...	3-15 Years	Nitsuchiba	James Sushi	Japan	Elvis Presley	C	0	0	0
11	342.400	98.322	Men´s Clothes	Fashion for ...	Sunny Clothes	Peter Shilton	UK	Lennart Sko...	C	0	0	1
12	320.950	-33.742	Sportwear	Sports...	New Balls	Pat Cash	Australia	Lennart Sko...	B	0	0	1
13	1077.250	161.001	Sportwear	Sports...	Der Bahnhof	Karl Strauba...	Germany	Helen Brolin	B	1	0	0
14	192.600	1085.695	Womens wear	Fashion for ...	Austerlich	Karl Heinz B...	Germany	Helen Brolin	B	1	0	0
15	478.140	302.391	Men´s Clothes	Fashion for ...	Big L	James Brown	USA	Tom Lindwall	C	1	0	0
16	4.370	363.947	Babywear	0-3 Years	ABC	Anna Book	Sweden	Rob Carsson	C	1	0	0
17	540.380	937.374	Sportwear	Sports...	Fast Runners	Ben Johnsson	Canada	Leif Shine	B	0	1	0



Saving the model predictions for further usage. Deploying the best model:



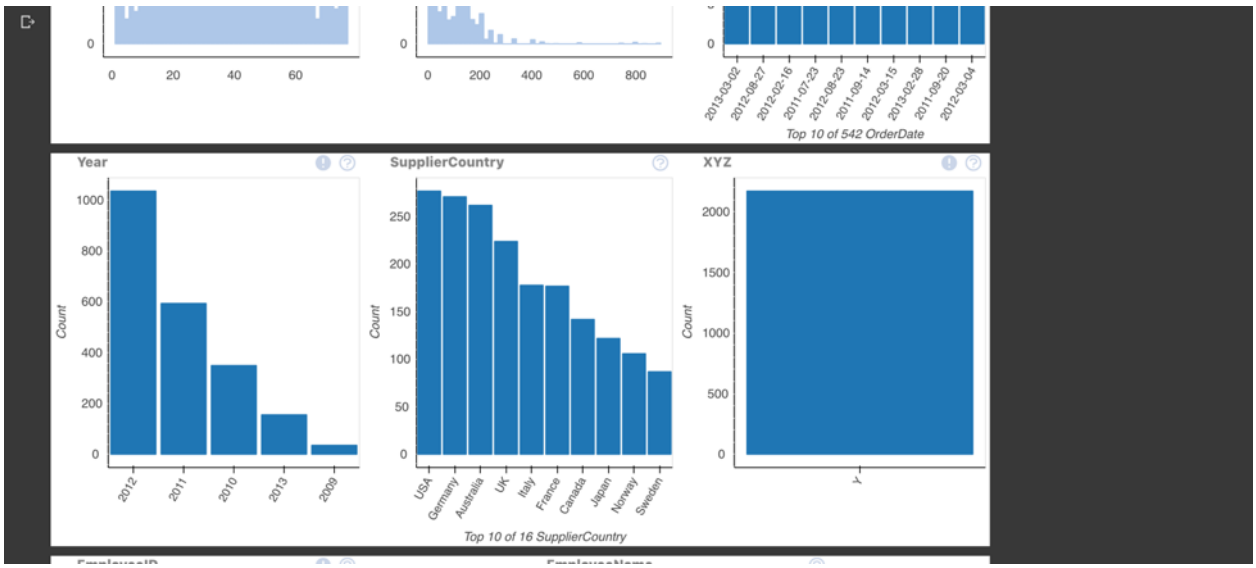
This model shows quite good results, but let's check what models we can create using other tools as well.

5. Data Analysis and building the model using Python in Colab with loading the data from Google BigQuery

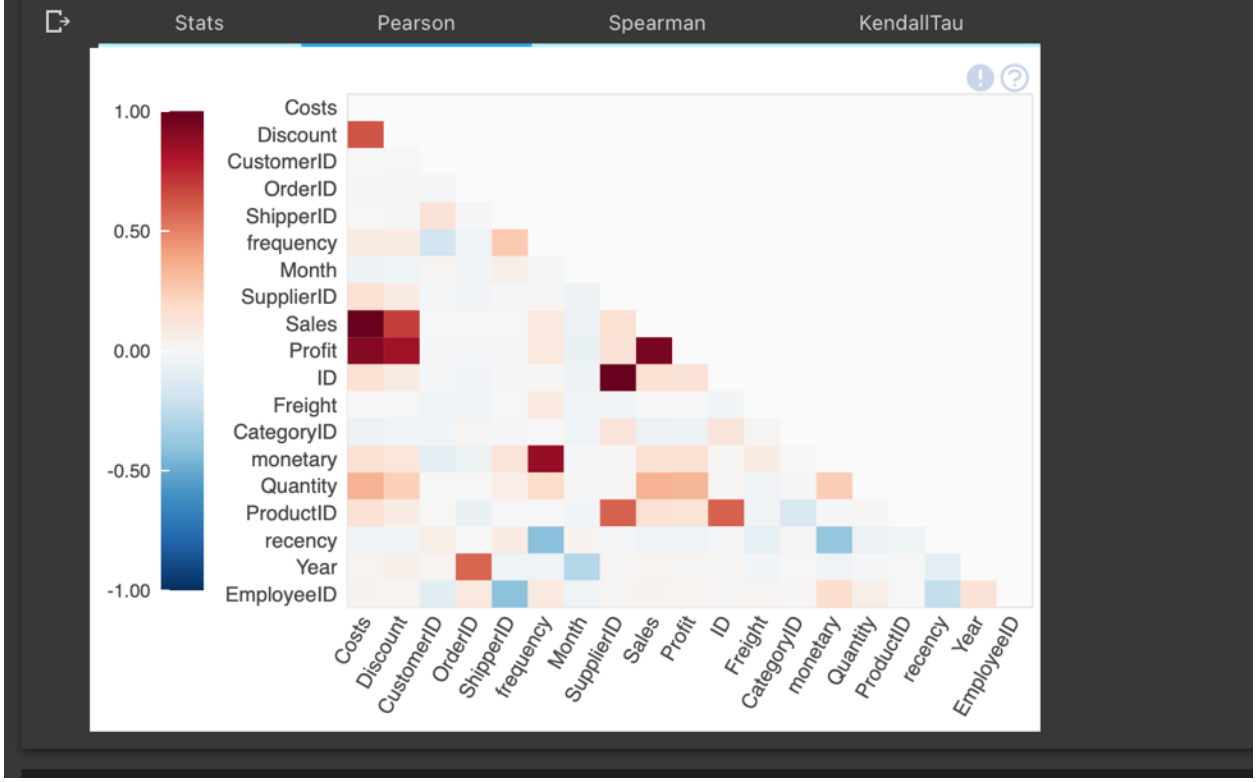
We will use the CSV file enriched with **ABC, XYZ, RFM analysis**, the file is loaded from **Google BigQuery** to **Colab**. The link to **Colab Notebook** is provided below.
https://colab.research.google.com/drive/1qDyCp2jGx02AQbtUxz_FoQRSxGueoDQV?usp=sharing

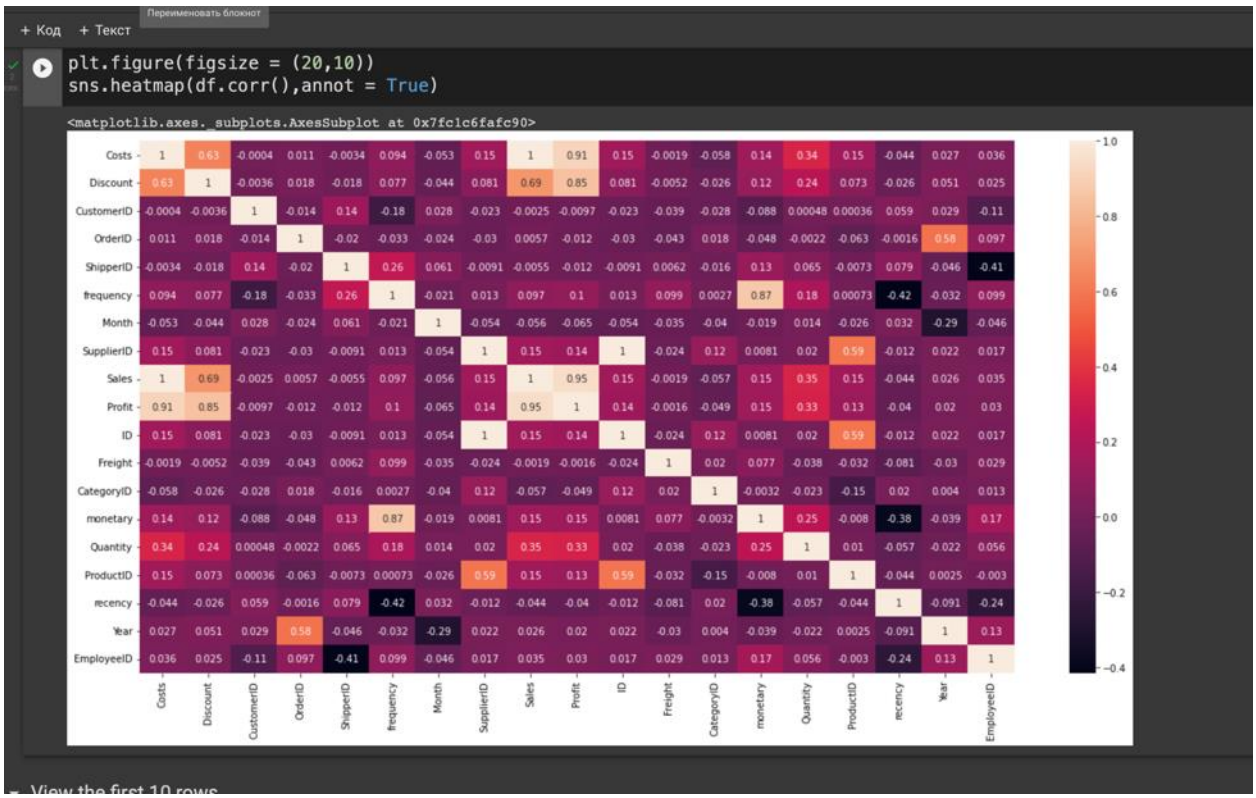
5.1. EDA

We performed **EDA** in **Colab notebook** using **Pandas Profiling, Dataprep** libraries. **Dataprep** helped us to understand that there are no missing values in the dataset, there are no repeated lines in the dataset.



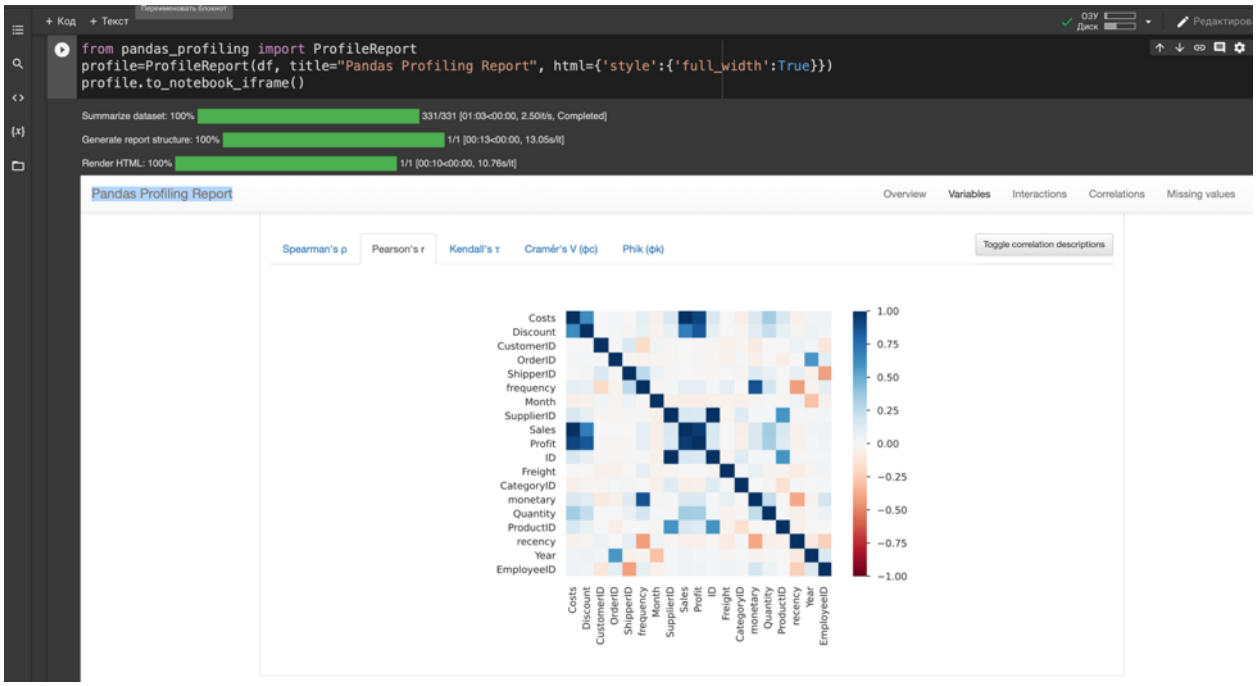

```
#Checking correlation
plot_correlation(df)
```

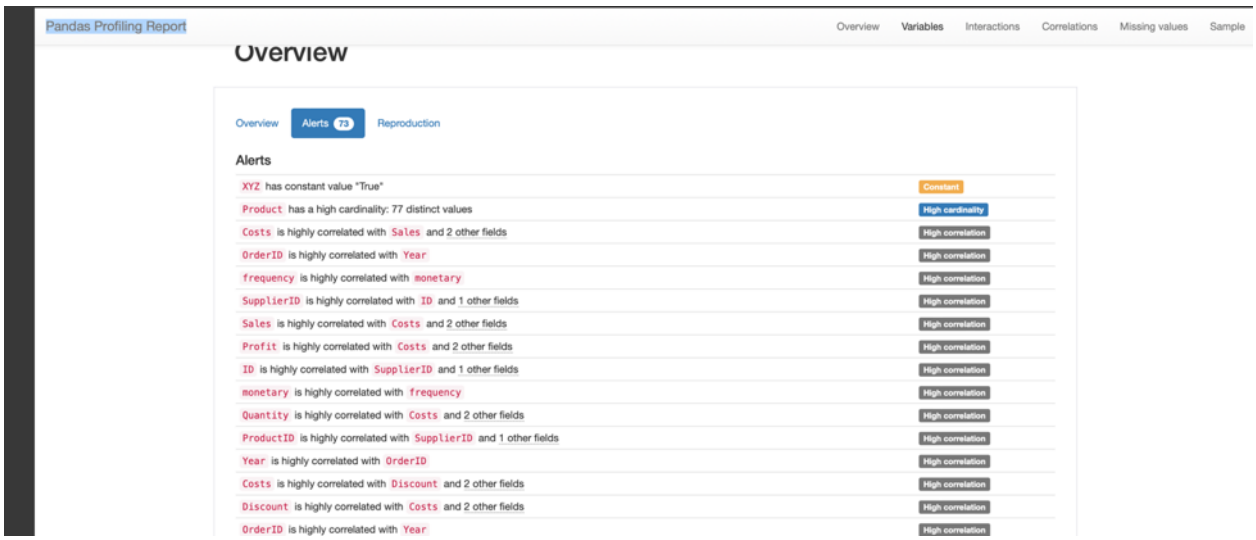




Using **Pandas Profiling** library, we are getting more detailed information on the dataset. The report is saved in the file below.

https://drive.google.com/file/d/1_CH-cDYTNO3a5kvSSeCH_2LuCjStkPu9/view?usp=sharing





5.2. Conclusions

1. **ABC analysis** has shown that there are only items from **categories B and C** present in our dataset.
2. **XYZ analysis** has shown that there are only **category Y** items present in our shop.
3. The **discount value** in 92% of orders is **between USD 0 and USD 146.**
4. There are **8 categories** of items being sold:
    - **Sportswear** is **26%** of all items in the shop;
    - **Kid's clothes, men's and women's clothes** are **14%** each;
    - The rest of the categories have less than 10% each.

5. The largest bin for **transaction values** is for the bin with translations varying from **USD 2 to USD 877**, we can observe it in **87%** of all sales cases. The next bin for sales is for transaction values varying between **USD 877 and USD 1752**, this bin represents **13%** of all cases.
6. **Profit** for our shop lies **between 0 and USD 232** in majority of cases **(84%).**
7. We can see that the majority of sales happened in 2012, **48%** of all sales are date back to **2012**. It would be interesting to check if the sales values correlate to the amount of transaction in 2012 and if the sales in monetary value are also the highest in 2012.
8. **Main suppliers** for the goods presented in our shop are **USA, Germany, Australia, UK**.

We can see high correlation between cost and discount, also between profit and discount. These features are interconnected, so we did not receive any new information from this correlation, we will remove these features when building the model.

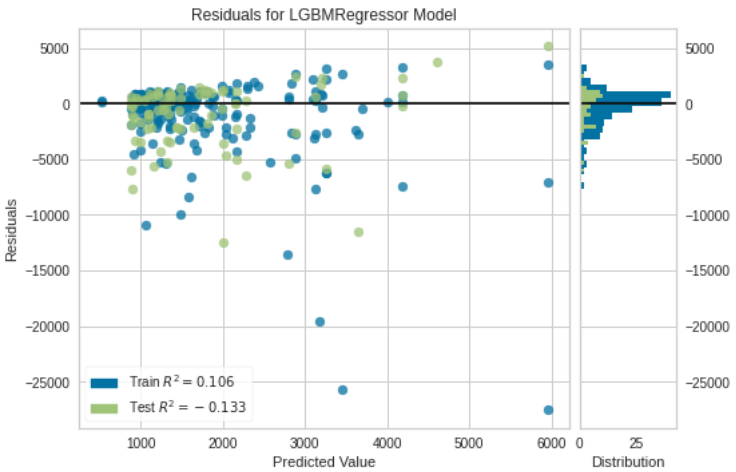
5.3. Building the model with PyCaret

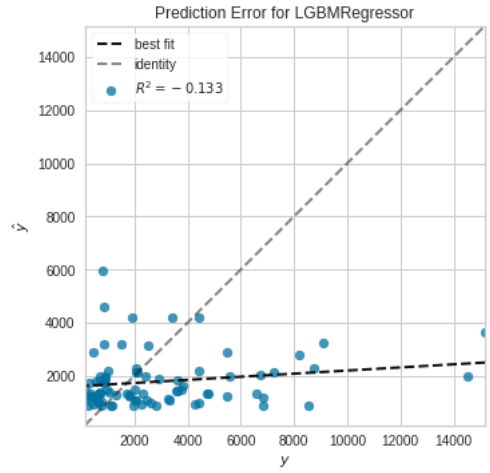
1 to 18 of 18 entriesFilter

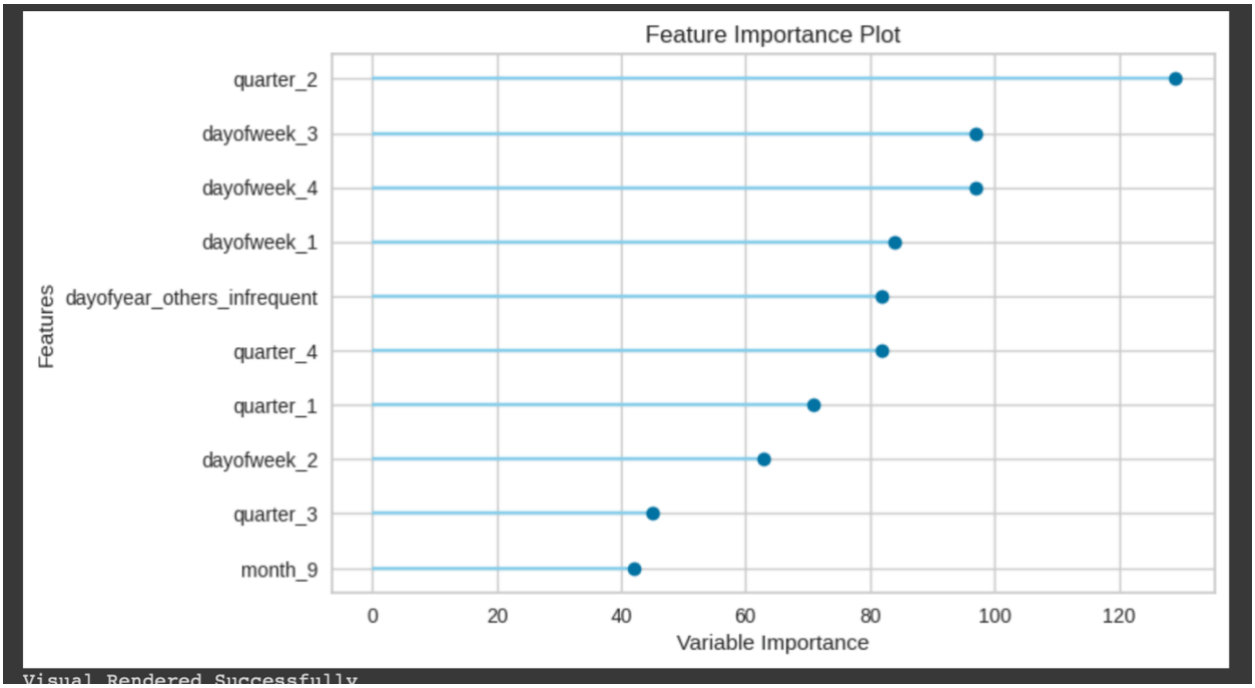
index	Model	MAE	
lightgbm	Light Gradient Boosting Machine	2268.846	184963
br	Bayesian Ridge	2202.2734	196582
lasso	Lasso Regression	2204.0901	198569
dummy	Dummy Regressor	2204.0901	198569
llar	Lasso Least Angle Regression	2204.0901	198569
en	Elastic Net	2204.0901	198569
ada	AdaBoost Regressor	2212.332	196875
knn	K Neighbors Regressor	2319.8667	192567
rf	Random Forest Regressor	2401.0834	205689
gbr	Gradient Boosting Regressor	2524.2977	216003
ridge	Ridge Regression	2621.2648	224466
omp	Orthogonal Matching Pursuit	2677.3637	229154
et	Extra Trees Regressor	3363.5112	319007
dt	Decision Tree Regressor	3446.7288	354430
par	Passive Aggressive Regressor	4127.3513	843890
lr	Linear Regression	4396.1628	120461
huber	Huber Regressor	5321.4899	301565
lar	Least Angle Regression	6.546956456302107e+178	Infinity

We will take 3 best models and fine-tune hyperparameters.

The best model is **lightgbm**.







Parameters of the model:

```
PowerTransformedTargetRegressor(bagging_fraction=0.6, bagging_freq=3,
                                boosting_type='gbdt', class_weight=None,
                                colsample_bytree=1.0,
feature_fraction=0.8,
                                importance_type='split',
learning_rate=0.0005,
                                max_depth=-1, min_child_samples=11,
                                min_child_weight=0.001,
min_split_gain=0.1,
                                n_estimators=260, n_jobs=-1,
num_leaves=256,
                                objective=None,
                                power_transformer_metho...

importance_type='split',

learning_rate=0.0005,
                                                        max_depth=-1,
```

```
min_child_samples=11,

min_child_weight=0.001,

min_split_gain=0.1,
                                                    n_estimators=260,
                                                    n_jobs=-1,
                                                    num_leaves=256,
                                                    objective=None,
                                                    random_state=6026,
                                                    reg_alpha=0.3,
                                                    reg_lambda=0.4,
                                                    silent='warn',
                                                    subsample=1.0,

subsample_for_bin=200000,
                                                    subsample_freq=0),
                                  silent='warn', subsample=1.0,
                                  subsample_for_bin=200000,
subsample_freq=0)
_freq=0),
                                  silent='warn', subsample=1.0,
                                  subsample_for_bin=200000,
subsample_freq=0)
```

Metrics of **lightgbm** model created by **PyCaret** are shown below:

MAE	MSE	RMSE	R2
2043.7348	10223266.27	3197.3843	-0.1671

6. Tableau visualizations, dashboards and insights

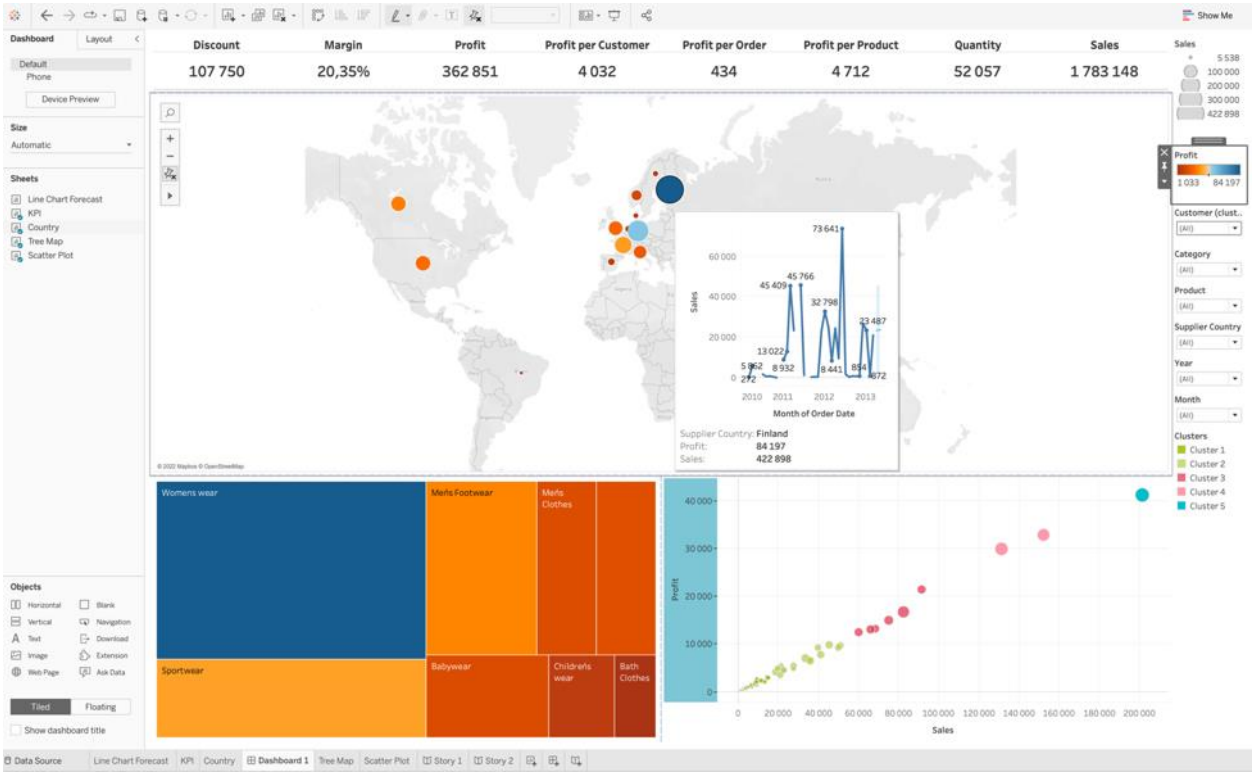
Visualizations were created using Tableau to provide some insights on the sales/profits, interactive dashboard containing main KPIs was created along with stories provide additional information to help in decision making.

Predictions for sales volumes created using Time Series in Tableau, clusterization of customers by categories done.

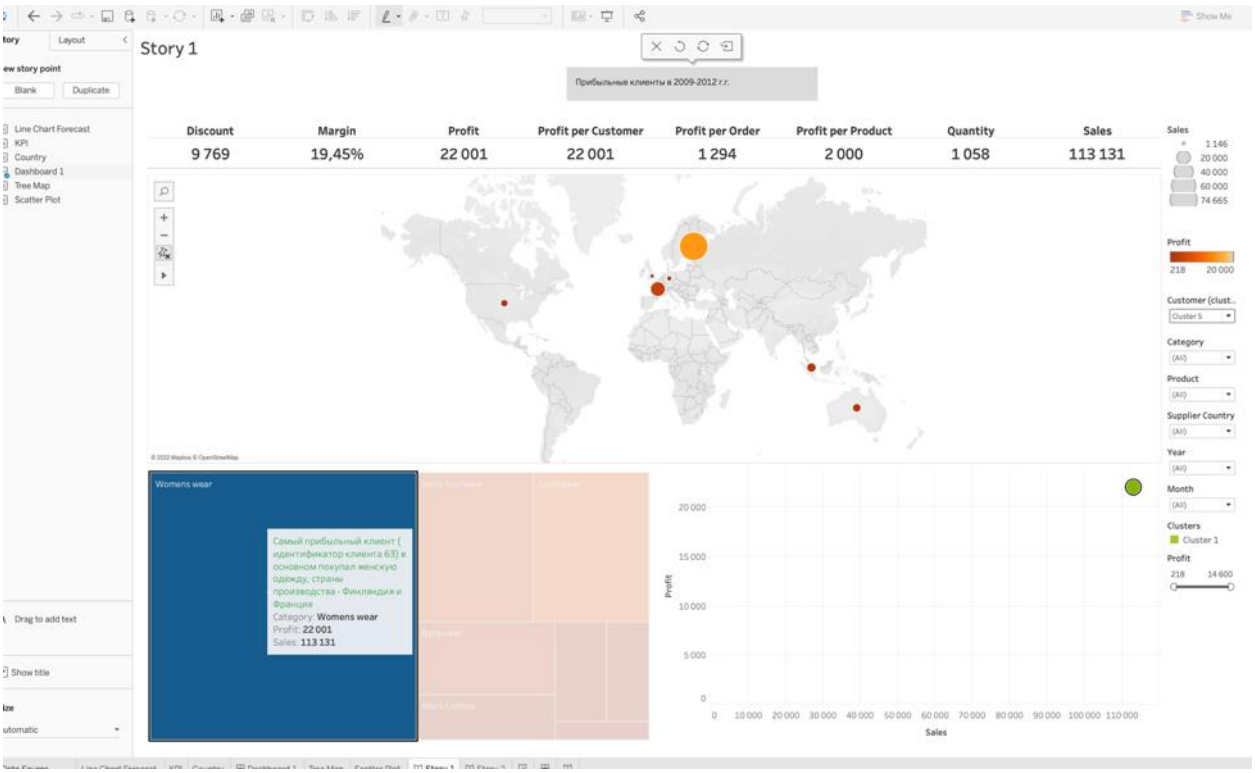
Links to interactive dashboard and stories are embedded in Colab Notebook:
https://colab.research.google.com/drive/1qDyCp2jGx02AQbtUxz_FoQRSxGueoDQV?usp=sharing#scrollTo=bxG2LPf_69cA

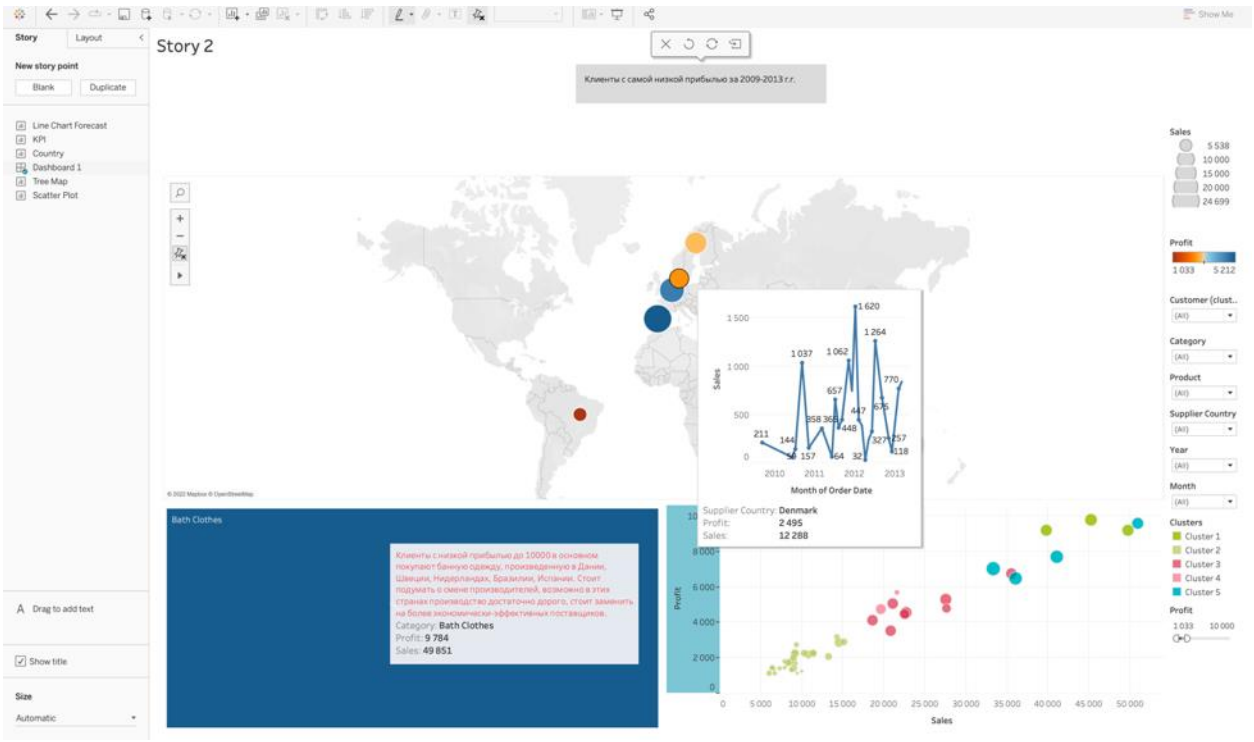
Links to the main dashboard and stories are also available below:
Customers with the Highest Profit
Customers with the Lowest Profit

The main KPIs of our shop are presented on the first dashboard. We have performed clusterization of the customers, there are 5 clusters based on profit value. Sales values, categories of goods that are sold in our shop, and suppliers' countries are presented on this dashboard.



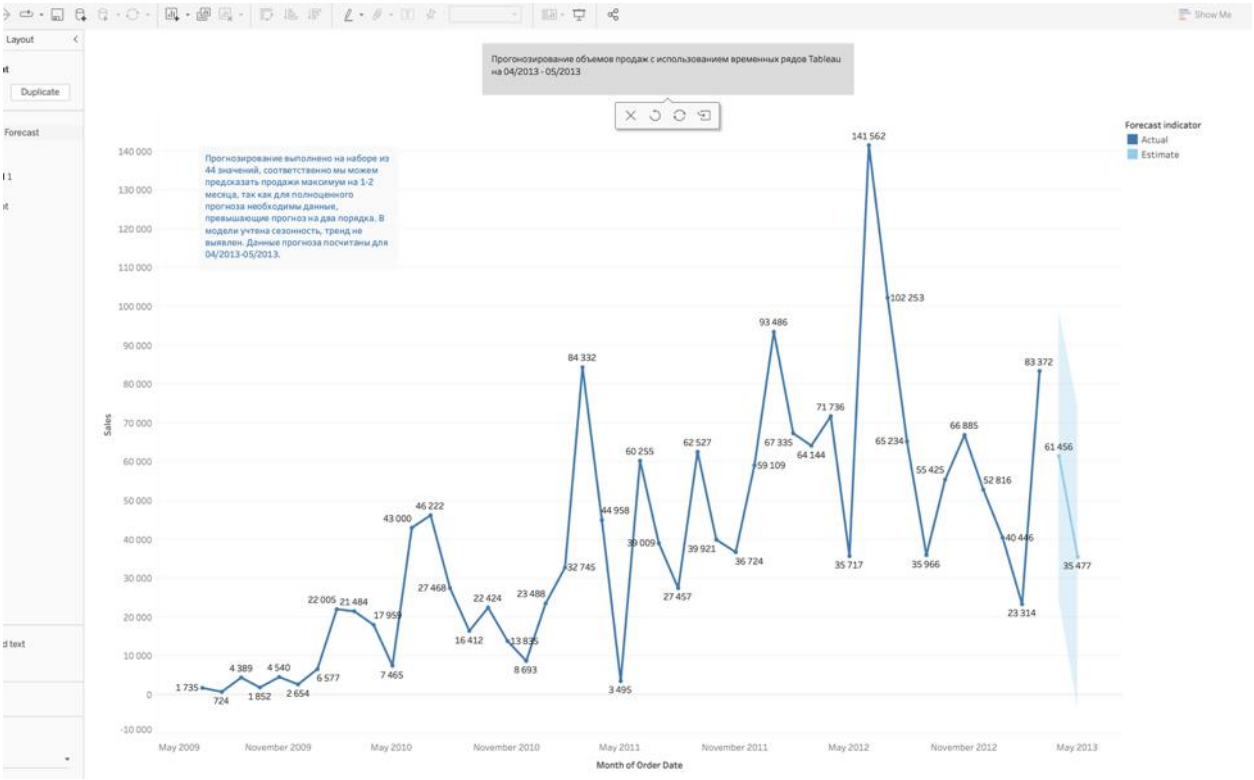
Insights on the clients with maximum and minimum profit were provided, these insights are presented in the Stories.





Predictive model to forecast **Sales Volumes** with **Time Series** was created in Tableau.

[Sales Forecast using Tableau TS]



Model's metrics for **Sales Volumes Prediction** are provided below. The metrics of the model created in Tableau are worse than the metrics of the model created using PyCaret in Python.

All forecasts were computed using exponential smoothing.

Sum of Sales

Model			Quality Metrics					Smoothing Coefficients		
Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Additive	Additive	Additive	19 022	12 935	0.62	76,5%	921	0.297	0.000	0.000

Options Used to Create Forecasts

Time series: Month of Order Date
Measures: Sum of Sales
Forecast forward: 2 months (April 2013 – May 2013)
Forecast based on: July 2009 – March 2013
Ignore last: No periods ignored
Seasonal pattern: 12 month cycle

Sum of Sales

Initial	Change From Initial	Seasonal Effect				Contribution		
April 2013	April 2013 – May 2013	High		Low		Trend	Season	Quality
61 456 ± 37 283	-25 979	March 2013	27 098	May 2013	-24 527	0,6%	99,4%	Ok

7. Creating the model to predict Sales Volumes using Google BigQuery

We are using the **Time Series** file (with aggregation by dates/sales volumes) that we exported to **Google Colab** in point 5. We are creating the model using **Time Series ARIMA**.

Case 3 A... ML | *Case 3 A... ATE | DISABLE EDITOR TABS

RUN | SAVE | SHARE | SCHEDULE | MORE

This query will process 2.87 KB when run.

```
1 SELECT * from
2 ML.ARIMA_EVALUATE (MODEL`my-project-dec9.Demo.Case3_TS_arima`)
3 Order by Variance DESC
```

Press Alt+F1 for Accessibility Options

Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	non_seasonal_p	non_seasonal_d	non_seasonal_q	has_drift	log_likelihood	AIC	variance	seasonal_periods	has_holiday_effect	has_spikes_an
1	0	1	0	false	-12508.644039309698	25019.288078619396	6815567.3778469115	[WEEKLY]	false	true
2	0	1	0	true	-12508.63134098632	25021.262681972639	6815438.8767887373	[WEEKLY]	false	true
3	1	1	0	false	-12501.934650416815	25007.86930083363	6747950.4687451459	[WEEKLY]	false	true
4	1	1	0	true	-12501.920309971207	25009.840619942413	6747806.7764323289	[WEEKLY]	false	true
5	0	1	1	false	-12496.6771338921	24997.3542677842	6695355.5751834167	[WEEKLY]	false	true
6	0	1	1	true	-12496.660327650559	24999.320655301119	6695190.0933928629	[WEEKLY]	false	true
7	2	1	0	false	-12472.294317902248	24950.588635804495	6456931.4546707962	[WEEKLY]	false	true
8	2	1	0	true	-12472.275680972494	24952.551361944988	6456752.7075718781	[WEEKLY]	false	true
9	3	1	0	false	-12470.421681063934	24948.843362127867	6438959.7779284324	[WEEKLY]	false	true
10	3	1	0	true	-12470.401706799496	24950.803413598991	6438768.715539258	[WEEKLY]	false	true
11	3	1	1	true	-12470.044251670299	24952.088503340598	6435341.9669306362	[WEEKLY]	false	true

Results per page: 50 | 1 – 42 of 42

PERSONAL HISTORY | PROJECT HISTORY | SAVED QUERIES | REFRESH

Case 3 A... ML | *Case 3 A... ATE | Case3 A... nts | DISABLE EDITOR TABS

RUN | SAVE | SHARE | SCHEDULE | MORE

This query will process 3.1 KB when run.

```
1 SELECT
2  *
3 FROM
4  ML.ARIMA_COEFFICIENTS(MODEL `my-project-dec9.Demo.Case3_TS_arima`)
```

Press Alt+F1 for Accessibility Options

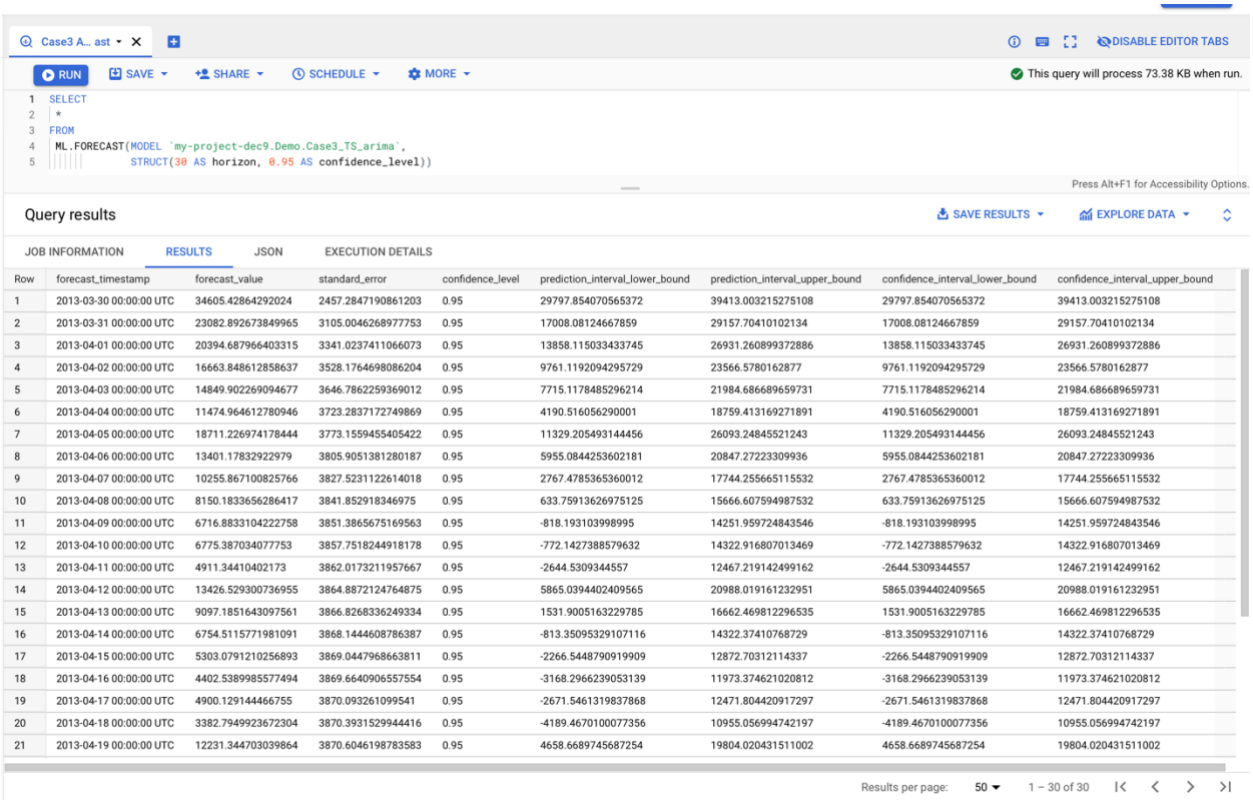
Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	ar_coefficients	ma_coefficients	intercept_or_drift
1	[0.8122055782756058]	Row: 1, ma_coefficients: -1.0404867591123528 Row: 2, ma_coefficients: -0.085630428671113493 Row: 3, ma_coefficients: 0.179246397753438 Row: 4, ma_coefficients: -0.053115924556110014	2.4446049957383633

PERSONAL HISTORY | PROJECT HISTORY | SAVED QUERIES | REFRESH



Case3 A... ast

DISABLE EDITOR TABS

RUN | SAVE | SHARE | SCHEDULE | MORE

This query will process 73.38 KB when run.

```
SELECT
  *
FROM
  ML.FORECAST(MODEL `my-project-dec9.Demo.Case3_TS_arima`,
              STRUCT(30 AS horizon, 0.95 AS confidence_level))
```

Press Alt+F1 for Accessibility Options.

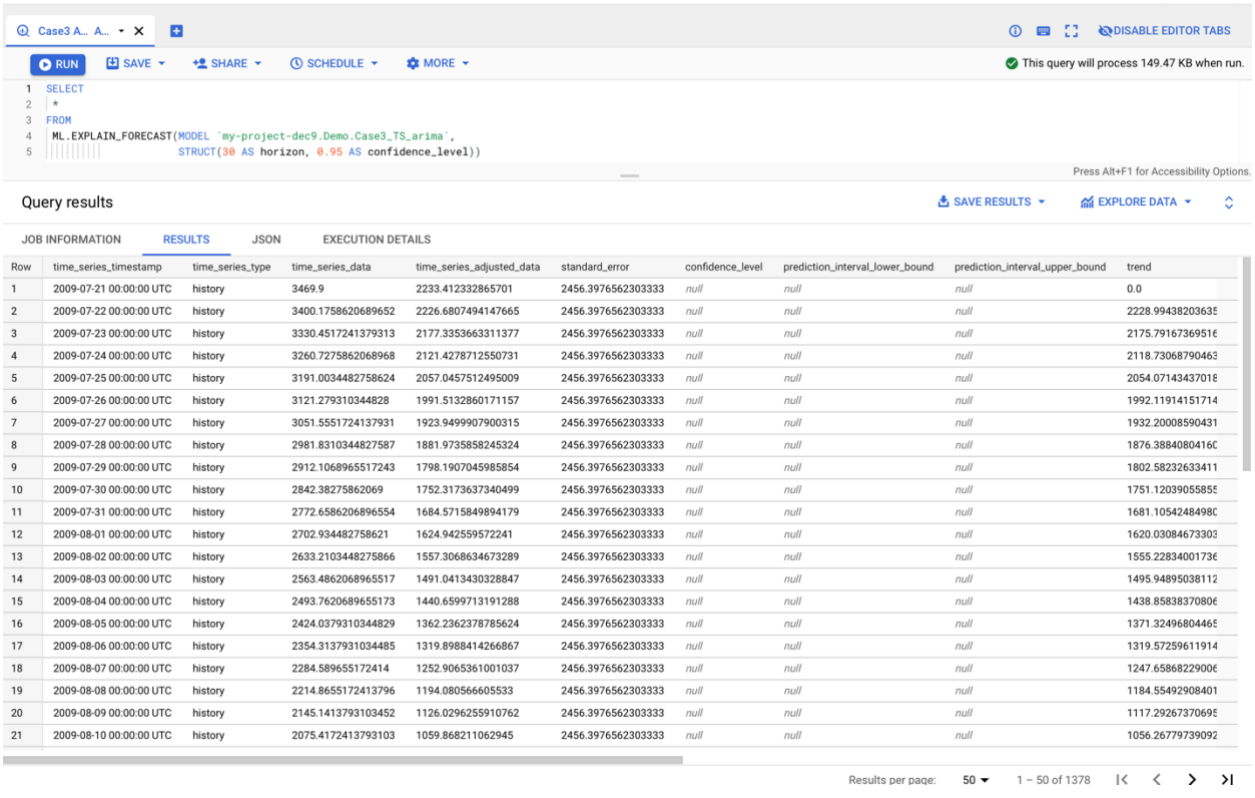
Query results

SAVE RESULTS | EXPLORE DATA

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	forecast_timestamp	forecast_value	standard_error	confidence_level	prediction_interval_lower_bound	prediction_interval_upper_bound	confidence_interval_lower_bound	confidence_interval_upper_bound
1	2013-03-30 00:00:00 UTC	34605.42864292024	2457.2847190861203	0.95	29797.854070565372	39413.003215275108	29797.854070565372	39413.003215275108
2	2013-03-31 00:00:00 UTC	23082.892673849965	3105.0046268977753	0.95	17008.08124667859	29157.70410102134	17008.08124667859	29157.70410102134
3	2013-04-01 00:00:00 UTC	20394.687966403315	3341.0237411066073	0.95	13858.115033433745	26931.260899372886	13858.115033433745	26931.260899372886
4	2013-04-02 00:00:00 UTC	16663.848612858637	3528.1764698086204	0.95	9761.1192094295729	23566.5780162877	9761.1192094295729	23566.5780162877
5	2013-04-03 00:00:00 UTC	14849.902269094677	3646.7862259369012	0.95	7715.1178485296214	21984.686689659731	7715.1178485296214	21984.686689659731
6	2013-04-04 00:00:00 UTC	11474.964612780946	3723.2837172749869	0.95	4190.516056290001	18759.413169271891	4190.516056290001	18759.413169271891
7	2013-04-05 00:00:00 UTC	18711.226974178444	3773.1559455405422	0.95	11329.205493144456	26093.24845521243	11329.205493144456	26093.24845521243
8	2013-04-06 00:00:00 UTC	13401.17832922979	3805.9051381280187	0.95	5955.0844253602181	20847.27223309936	5955.0844253602181	20847.27223309936
9	2013-04-07 00:00:00 UTC	10255.867100825766	3827.5231122614018	0.95	2767.4785365360012	17744.255665115532	2767.4785365360012	17744.255665115532
10	2013-04-08 00:00:00 UTC	8150.1833656286417	3841.852918346975	0.95	633.75913626975125	15666.607594987532	633.75913626975125	15666.607594987532
11	2013-04-09 00:00:00 UTC	6716.8833104222758	3851.3865675169563	0.95	-818.193103998995	14251.959724843546	-818.193103998995	14251.959724843546
12	2013-04-10 00:00:00 UTC	6775.387034077753	3857.7518244918178	0.95	-772.1427388579632	14322.916807013469	-772.1427388579632	14322.916807013469
13	2013-04-11 00:00:00 UTC	4911.34410402173	3862.0173211957667	0.95	-2644.5309344557	12467.219142499162	-2644.5309344557	12467.219142499162
14	2013-04-12 00:00:00 UTC	13426.529300736955	3864.8872124764875	0.95	5865.0394402409565	20988.019161232951	5865.0394402409565	20988.019161232951
15	2013-04-13 00:00:00 UTC	9097.1851643097561	3866.8268336249334	0.95	1531.9005163229785	16662.469812296535	1531.9005163229785	16662.469812296535
16	2013-04-14 00:00:00 UTC	6754.5115771981091	3868.1444608786387	0.95	-813.35095329107116	14322.37410768729	-813.35095329107116	14322.37410768729
17	2013-04-15 00:00:00 UTC	5303.0791210256893	3869.0447968663811	0.95	-2266.5448790919909	12872.70312114337	-2266.5448790919909	12872.70312114337
18	2013-04-16 00:00:00 UTC	4402.5389985577494	3869.6640906557554	0.95	-3168.2966239053139	11973.374621020812	-3168.2966239053139	11973.374621020812
19	2013-04-17 00:00:00 UTC	4900.129144466755	3870.093261099541	0.95	-2671.5461319837868	12471.80442017297	-2671.5461319837868	12471.80442017297
20	2013-04-18 00:00:00 UTC	3382.7949923672304	3870.3931529944416	0.95	-4189.4670100077356	10955.056994742197	-4189.4670100077356	10955.056994742197
21	2013-04-19 00:00:00 UTC	12231.344703039864	3870.6046198783583	0.95	4658.6689745687254	19804.020431511002	4658.6689745687254	19804.020431511002

Results per page: 50 | 1 – 30 of 30



Case3 A... A...

DISABLE EDITOR TABS

RUN | SAVE | SHARE | SCHEDULE | MORE

This query will process 149.47 KB when run.

```
SELECT
  *
FROM
  ML.EXPLAIN_FORECAST(MODEL `my-project-dec9.Demo.Case3_TS_arima`,
                      STRUCT(30 AS horizon, 0.95 AS confidence_level))
```

Press Alt+F1 for Accessibility Options.

Query results

SAVE RESULTS | EXPLORE DATA

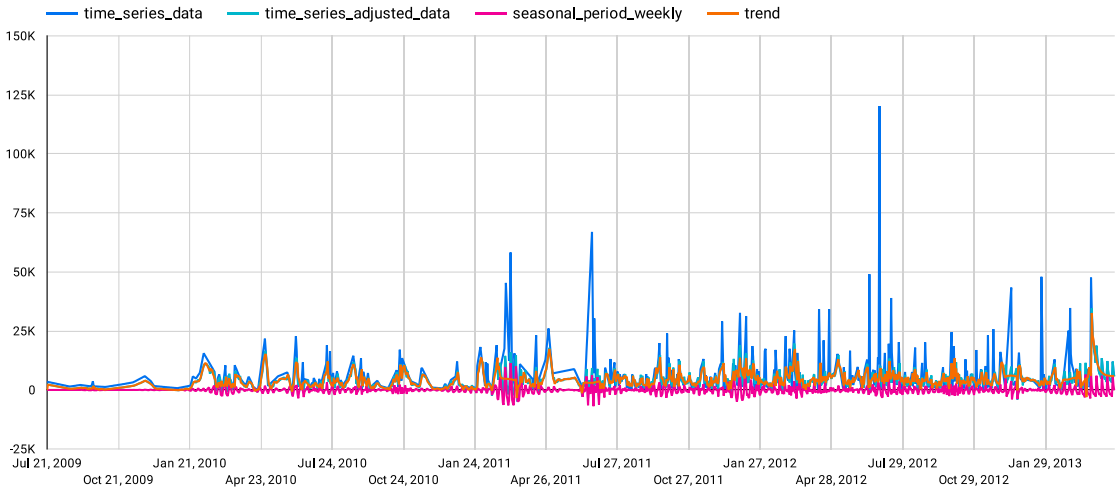
JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

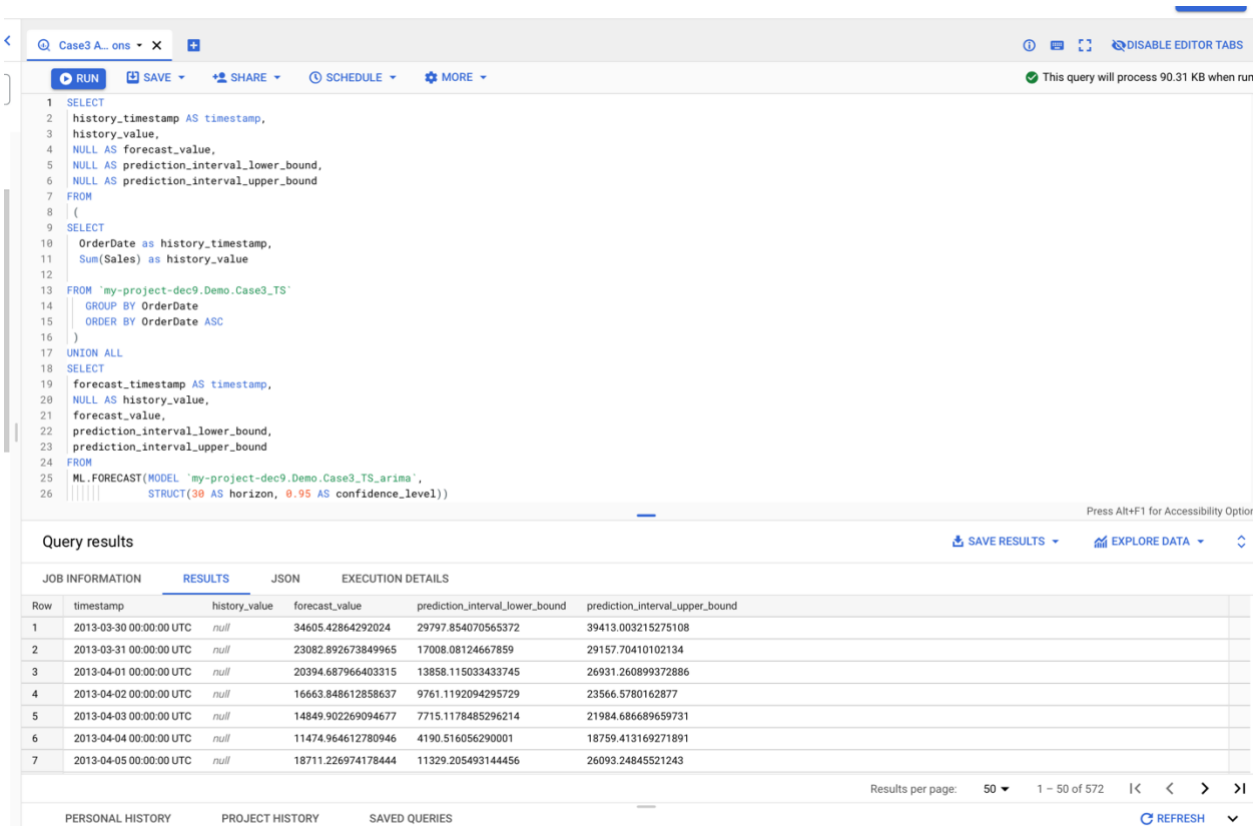
Row	time_series_timestamp	time_series_type	time_series_data	time_series_adjusted_data	standard_error	confidence_level	prediction_interval_lower_bound	prediction_interval_upper_bound	trend
1	2009-07-21 00:00:00 UTC	history	3469.9	2233.412332865701	2456.3976562303333	null	null	null	0.0
2	2009-07-22 00:00:00 UTC	history	3400.1758620689652	2226.6807494147665	2456.3976562303333	null	null	null	2228.994382036
3	2009-07-23 00:00:00 UTC	history	3330.4517241379313	2177.3353663311377	2456.3976562303333	null	null	null	2175.791673695
4	2009-07-24 00:00:00 UTC	history	3260.7275862068968	2121.4278712550731	2456.3976562303333	null	null	null	2118.730678904
5	2009-07-25 00:00:00 UTC	history	3191.0034482758624	2057.0457512495009	2456.3976562303333	null	null	null	2054.071434370
6	2009-07-26 00:00:00 UTC	history	3121.279310344828	1991.5132860171157	2456.3976562303333	null	null	null	1992.119141517
7	2009-07-27 00:00:00 UTC	history	3051.5551724137931	1923.9499907900315	2456.3976562303333	null	null	null	1932.200085904
8	2009-07-28 00:00:00 UTC	history	2981.8310344827587	1881.9735858245324	2456.3976562303333	null	null	null	1876.388408041
9	2009-07-29 00:00:00 UTC	history	2912.1068965517243	1798.1907045985854	2456.3976562303333	null	null	null	1802.582326334
10	2009-07-30 00:00:00 UTC	history	2842.38275862069	1752.3173637340499	2456.3976562303333	null	null	null	1751.120390558
11	2009-07-31 00:00:00 UTC	history	2772.6586206896554	1684.5715849894179	2456.3976562303333	null	null	null	1681.105424849
12	2009-08-01 00:00:00 UTC	history	2702.934482758621	1624.942559572241	2456.3976562303333	null	null	null	1620.030846733
13	2009-08-02 00:00:00 UTC	history	2633.2103448275866	1557.3068634673289	2456.3976562303333	null	null	null	1555.228340017
14	2009-08-03 00:00:00 UTC	history	2563.4862068965517	1491.0413430328847	2456.3976562303333	null	null	null	1495.948950381
15	2009-08-04 00:00:00 UTC	history	2493.7620689655173	1440.6599713191288	2456.3976562303333	null	null	null	1438.858383708
16	2009-08-05 00:00:00 UTC	history	2424.0379310344829	1362.2362378785624	2456.3976562303333	null	null	null	1371.324968046
17	2009-08-06 00:00:00 UTC	history	2354.3137931034485	1319.8988414266867	2456.3976562303333	null	null	null	1319.572596119
18	2009-08-07 00:00:00 UTC	history	2284.589655172414	1252.9065361001037	2456.3976562303333	null	null	null	1247.658682290
19	2009-08-08 00:00:00 UTC	history	2214.8655172413796	1194.080566605533	2456.3976562303333	null	null	null	1184.554929084
20	2009-08-09 00:00:00 UTC	history	2145.1413793103452	1126.0296255910762	2456.3976562303333	null	null	null	1117.292673706
21	2009-08-10 00:00:00 UTC	history	2075.4172413793103	1059.868211062945	2456.3976562303333	null	null	null	1056.267793909

Results per page: 50 | 1 – 50 of 1378

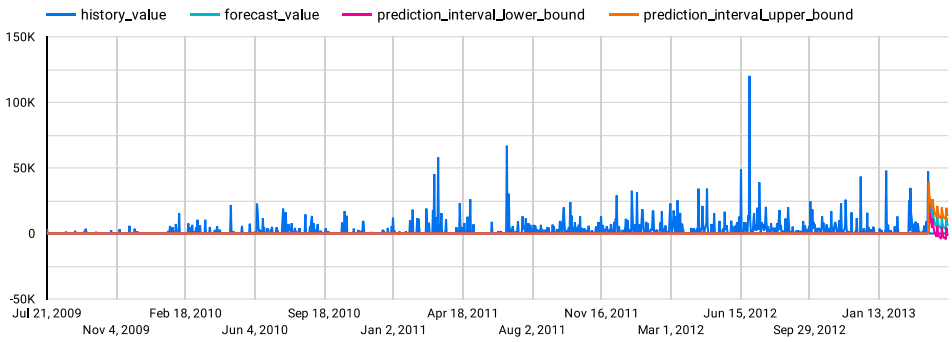
We visualize the forecast in **Google Data Studio**.

[Case 3 BigQuery ARIMA Forecast]





[BigQuery ARIMA Predictions in Data Studio]



8. Conclusion and comparing the models

Metrics of the model created in **BigQuery** are quite good. **ARIMA model** is the best model from the **3 models (ARIMA, Python PyCaret, Tableau) created with Time Series** to predict **Sales Volumes**.

The model created in **RapidMiner** to predict **Sales Volumes** uses not only statistical historic information (aggregated sales volumes/dates – **Time Series**), but also other features, so it makes sense to use both RapidMiner model and **ARIMA** model to get better predictions on **Sales Volumes.**