FAVORITA STORE SALES PREDICTION

AGENDA

- OVERVIEW
- DATASET AND PREPROCESSING
- EDA FINDINGS
- BASELINE MODELS
- NEXT STEPS

OVERVIEW

About the Business

Favorita is one of the **largest supermarket chains** in **Ecuador**, known for its extensive selection of **groceries, household items,** and other goods. It has **numerous locations** across the country.

Purpose & Impacts of the Project

- Given past sales data of different categories and stores, **predict future daily sales** of each category in different stores.
- Improve **customer experience**
- Control **business cost**

DATASET AND PREPROCESSING

5 Tables

3M rows, 12 Columns

1. Sales
2. Stores
3. Oil
4. Holidays & Events
5. Store Total Sales

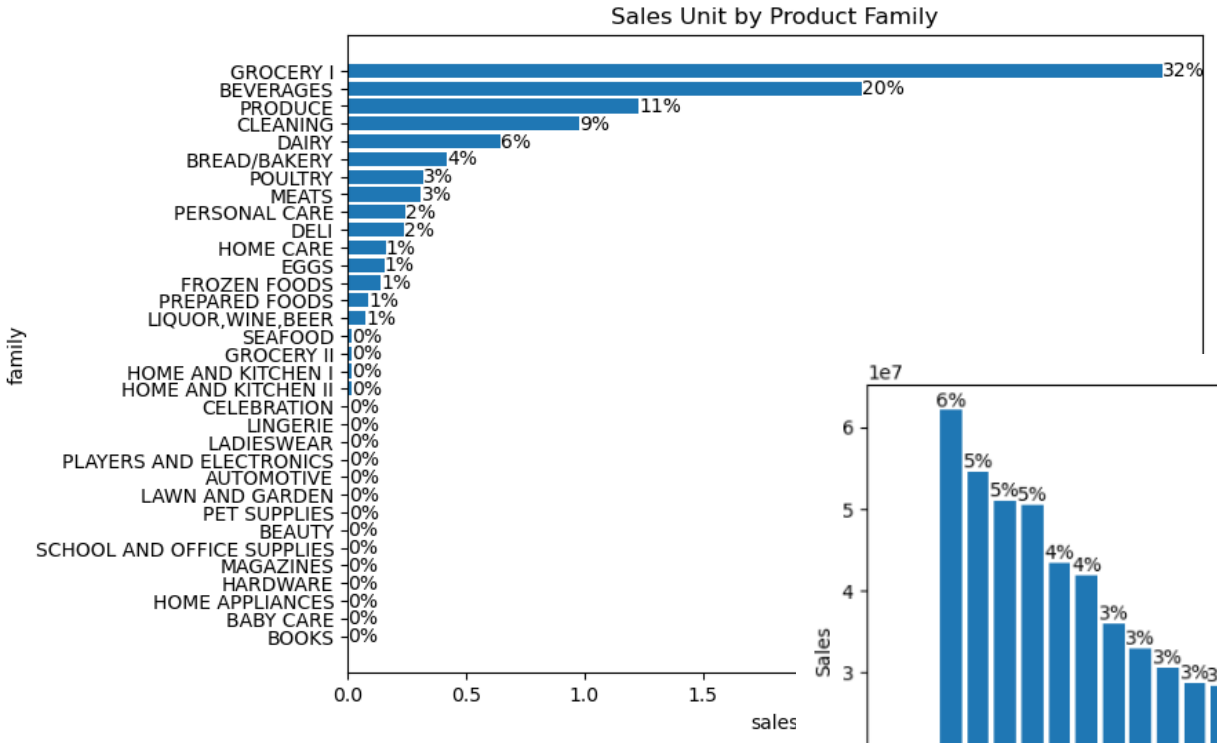
Preliminary Data Cleaning and EDA → In-depth Data Cleaning and EDA → Full Table EDA and Preprocessing

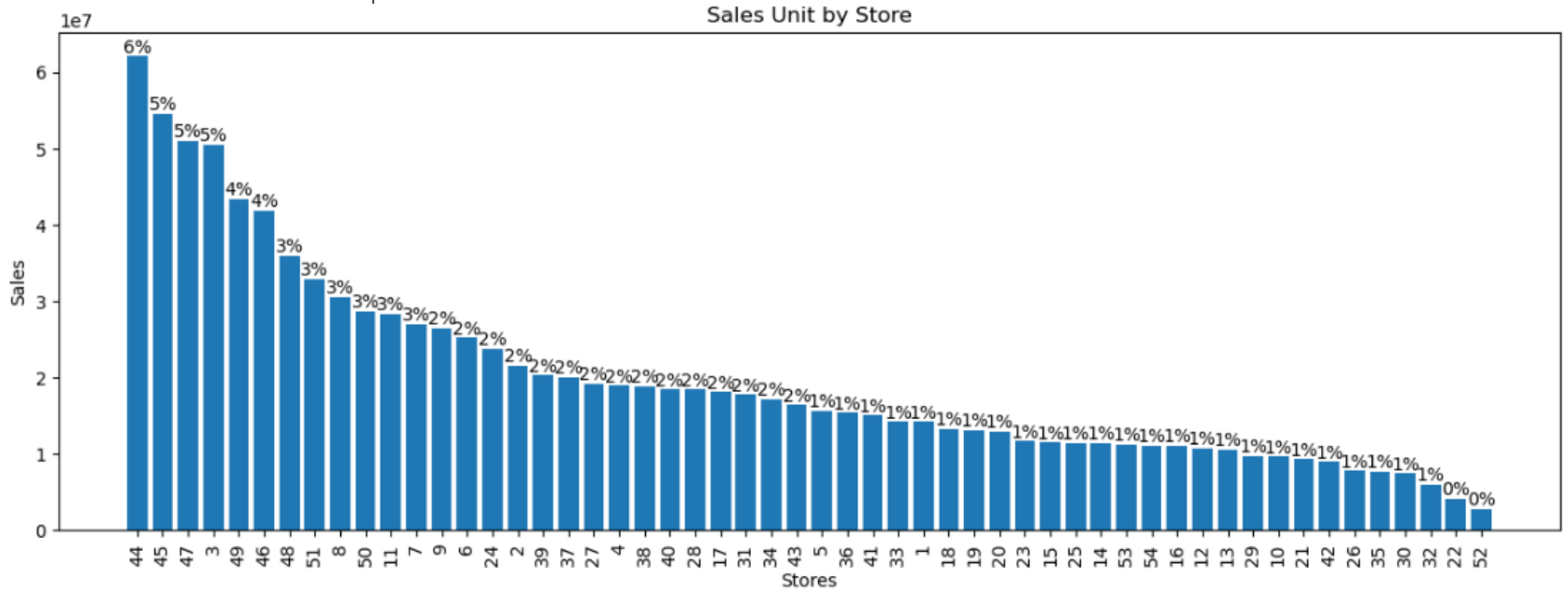
- Understand **each table's columns**
- Understand **the relationships** between each table
- Remove duplicates, change date format, locate 43 null values

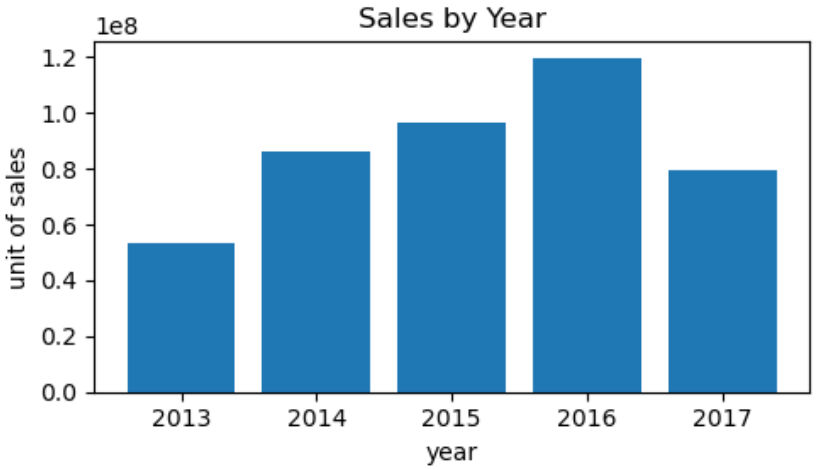
- **Define project scope:** focus on main families in the capital city
- Analyze **Xs' relationship with the target**
- Forward filled 43 null values and joined the five tables

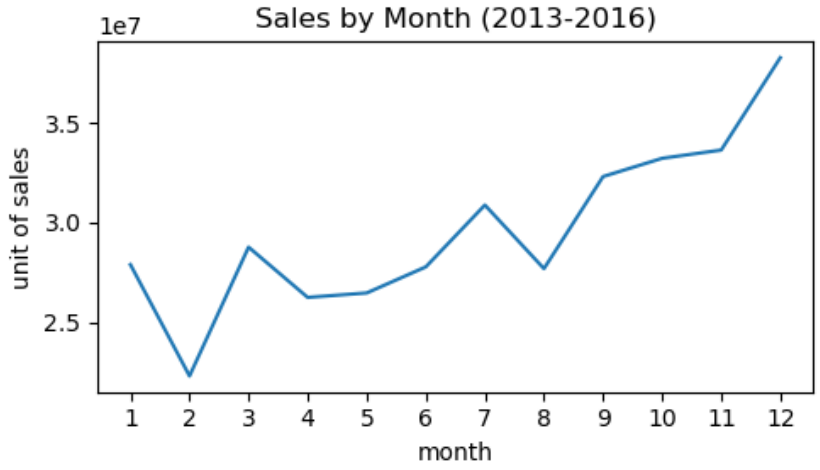
- Dropped **duplicated columns**
- Visualize and fill **nulls**
- Distribution of each column
- Translated the dataset
- Created dummy variables
- Time series column **feature engineering**

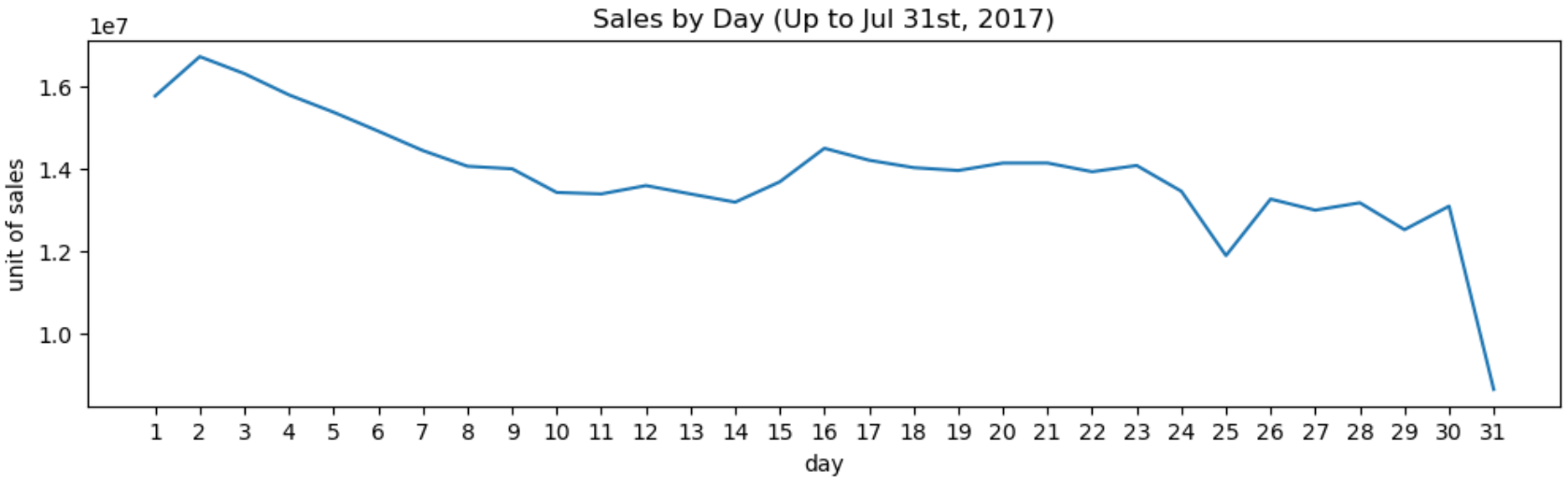
EDA: 51% of sales come from stores in Quito, and the top 3 categories account for 63% of total sales. The business heavily depends on Grocery I.







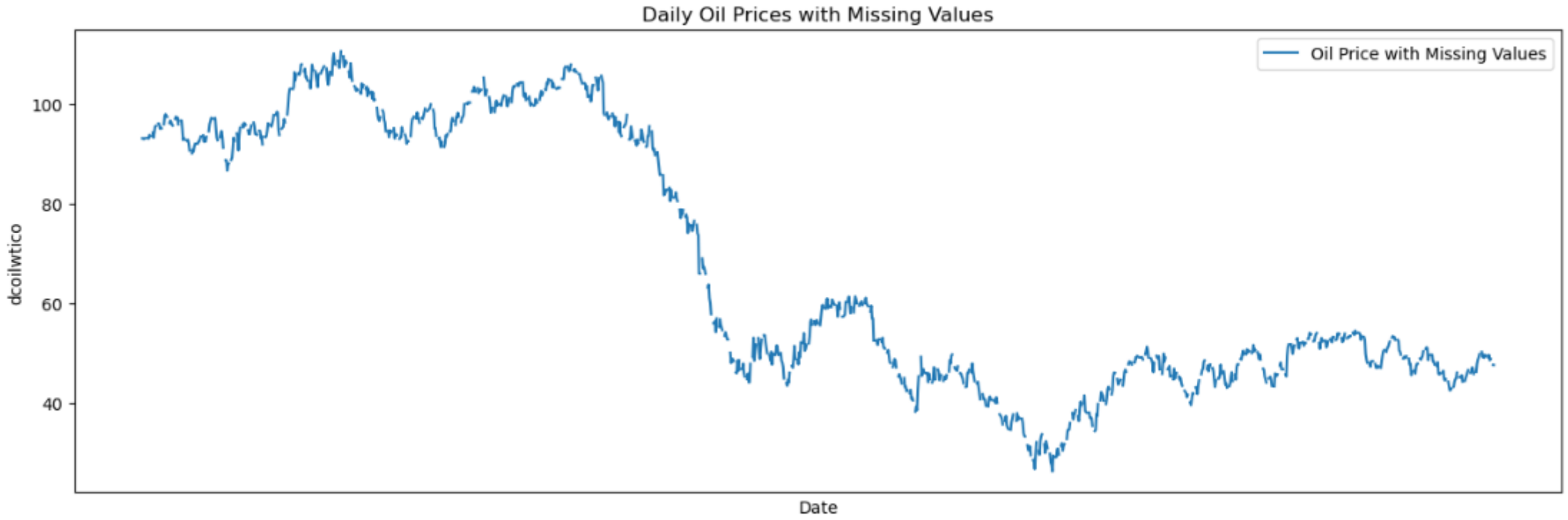




EDA:

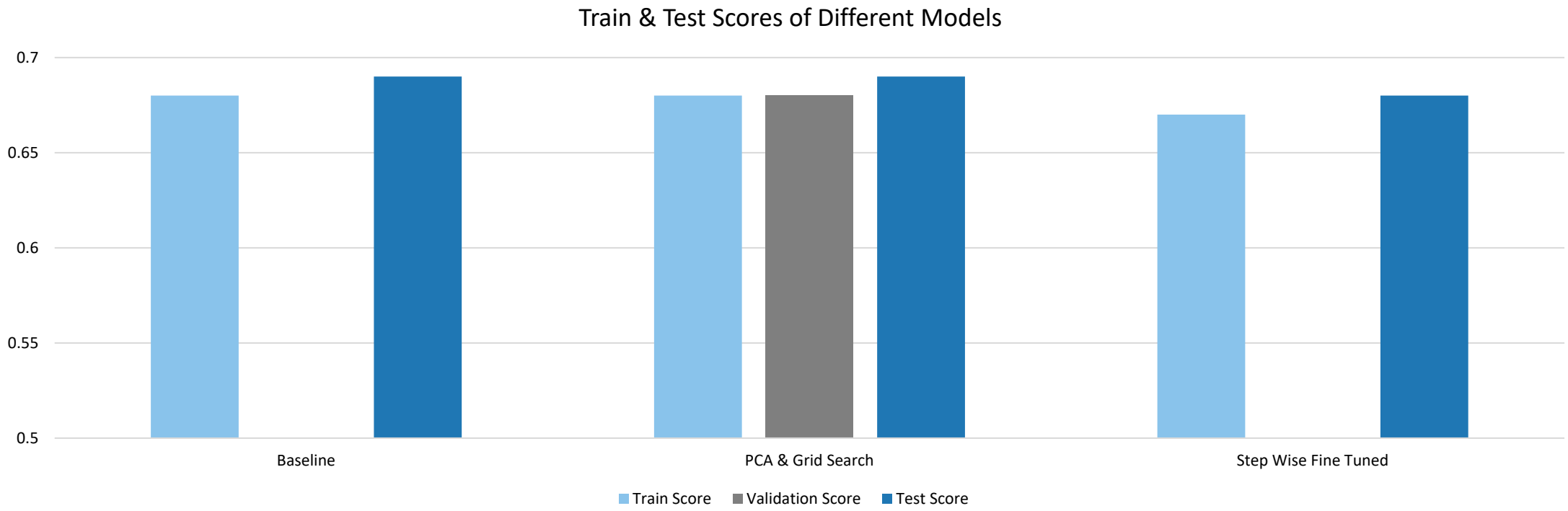
- From 2013 to 2016, sales increased steadily. The lower sales in 2017 are due to **incomplete data, only including up to July 31st**.
- Sales follow an upward trend monthly, peaking in December, with notable growth in March, July, and December.
- Sales decrease steadily during the first 14 days of the month, increase on day 15, then remain relatively stable until day 25.

EDA: The null values are **sparsely distributed** throughout the dataset. With **29%** of the data missing, forward or backward filling is unsuitable. Given the trends and seasonality in daily oil prices, a time series-specific interpolation method is used to fill the missing values.



BASELINE MODELS: ROBUST MODELS BUT LOW R2 SCORES

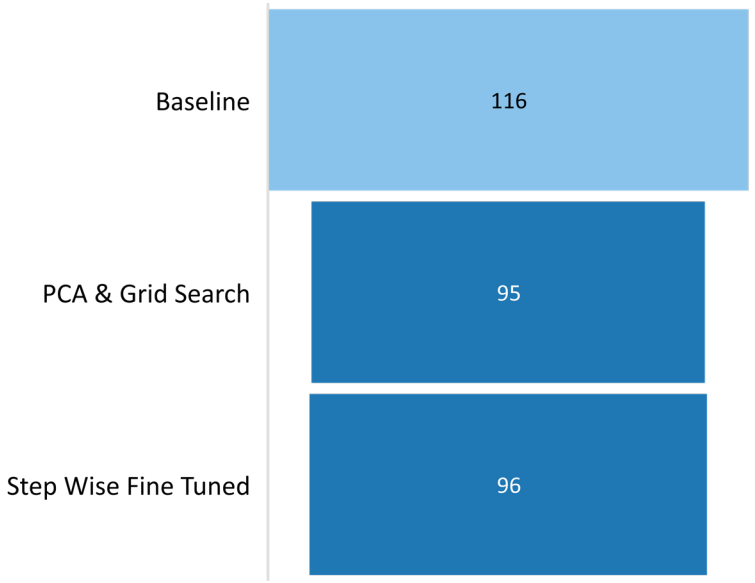
Baseline Model → **PCA Fine Tuning** → **Stepwise Fine Tuning**



MODEL EVALUATION: LINEARITY AND MULTICOLLINEARITY

- Only 20 features can be reduced to address multicollinearity.

Number of Features or PCs for Models

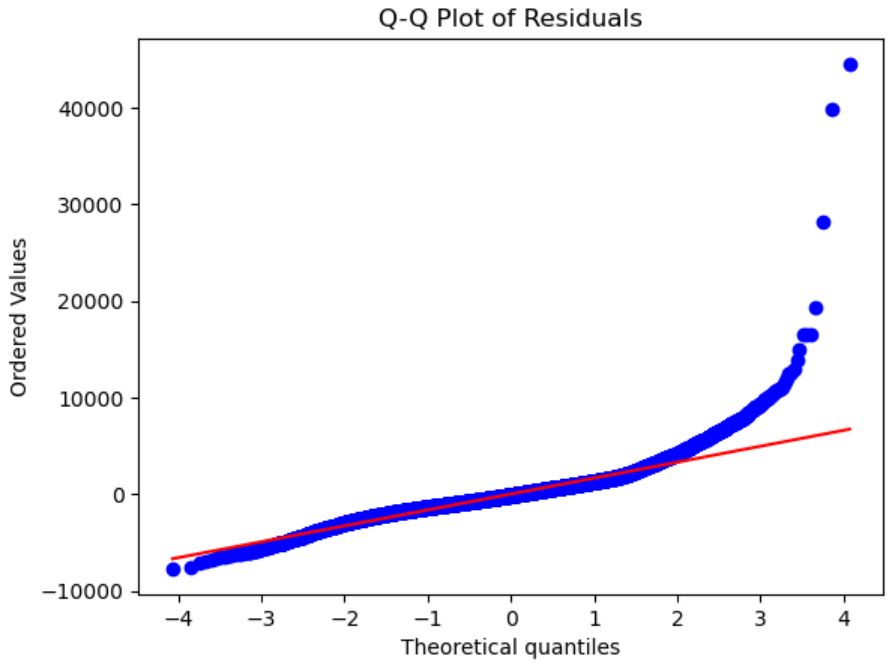


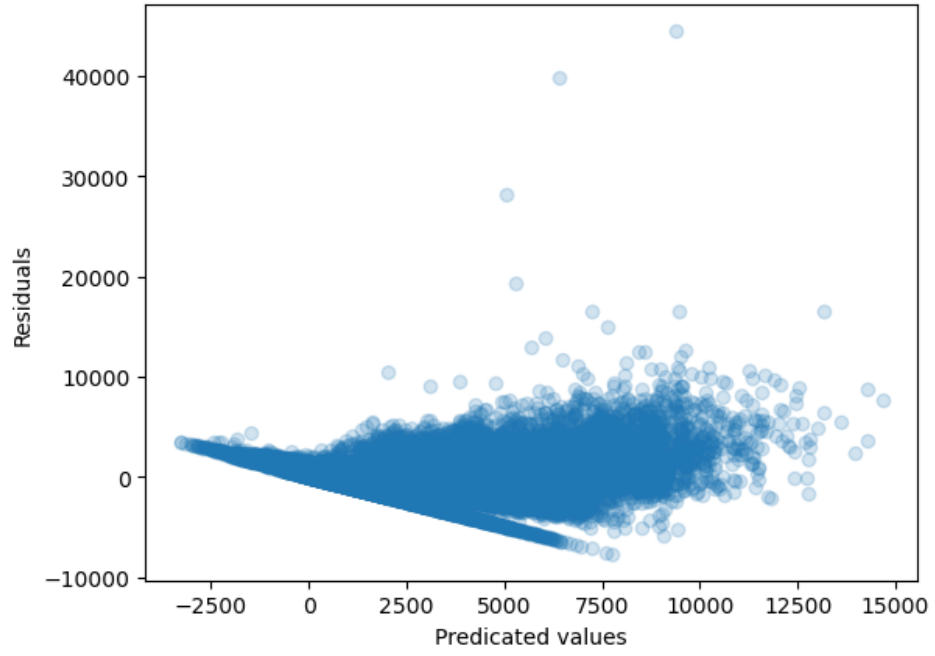
- Out of 116 features, only 24 show high correlations.
- Many features have low correlation with the target.

	feature1	feature2	feature_corr	feature1_y_corr	feature2_y_corr
0	sto_type_A	store_trans	0.73	0.35	0.52
1	cluster_11	store_nbr_45	0.68	0.19	0.17
2	cluster_11	store_nbr_49	0.69	0.19	0.09
3	cluster_12	store_nbr_17	1.00	-0.10	-0.10
4	cluster_15	store_nbr_10	1.00	-0.17	-0.17
5	cluster_5	store_nbr_44	1.00	0.23	0.23
6	cluster_6	store_nbr_9	0.79	-0.06	-0.04
7	cluster_6	sto_type_B	0.77	-0.06	-0.13
8	cluster_9	store_nbr_4	1.00	-0.10	-0.10
9	description_eng_Foundation of Quito-1	locale_Local	0.71	0.00	0.02
10	description_eng_foundation of Quito	locale_Local	0.70	0.02	0.02
11	year	dcoilwtico_interpolated	-0.83	0.24	-0.22
12	weekend	weekday	-0.75	0.18	-0.16

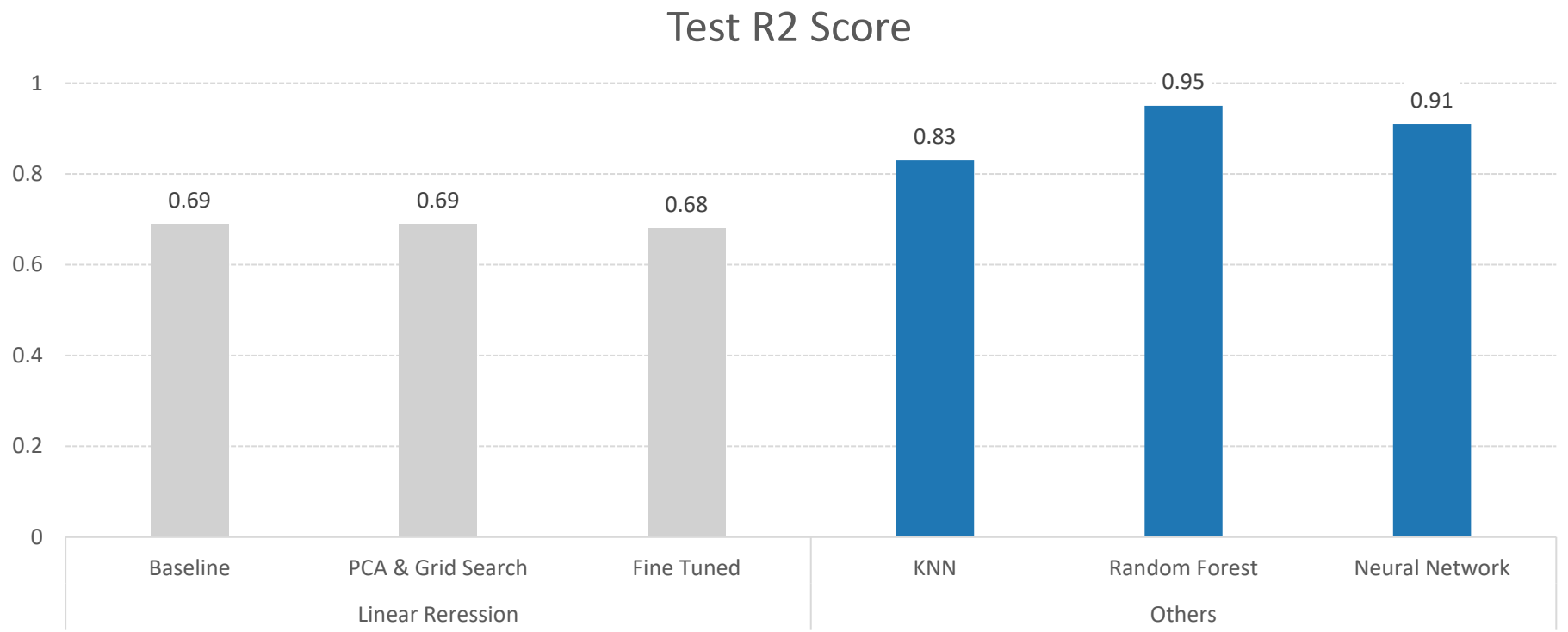
MODEL EVALUATION: RESIDUALS AND HOMOSCEDASTICITY

- Large gap at the tail & clear pattern seen between residuals and predicated values
- The model is not reliable





NEXT STEPS: TRY MORE ADVANCED AND NON-LINEAR MODELS




THANK YOU