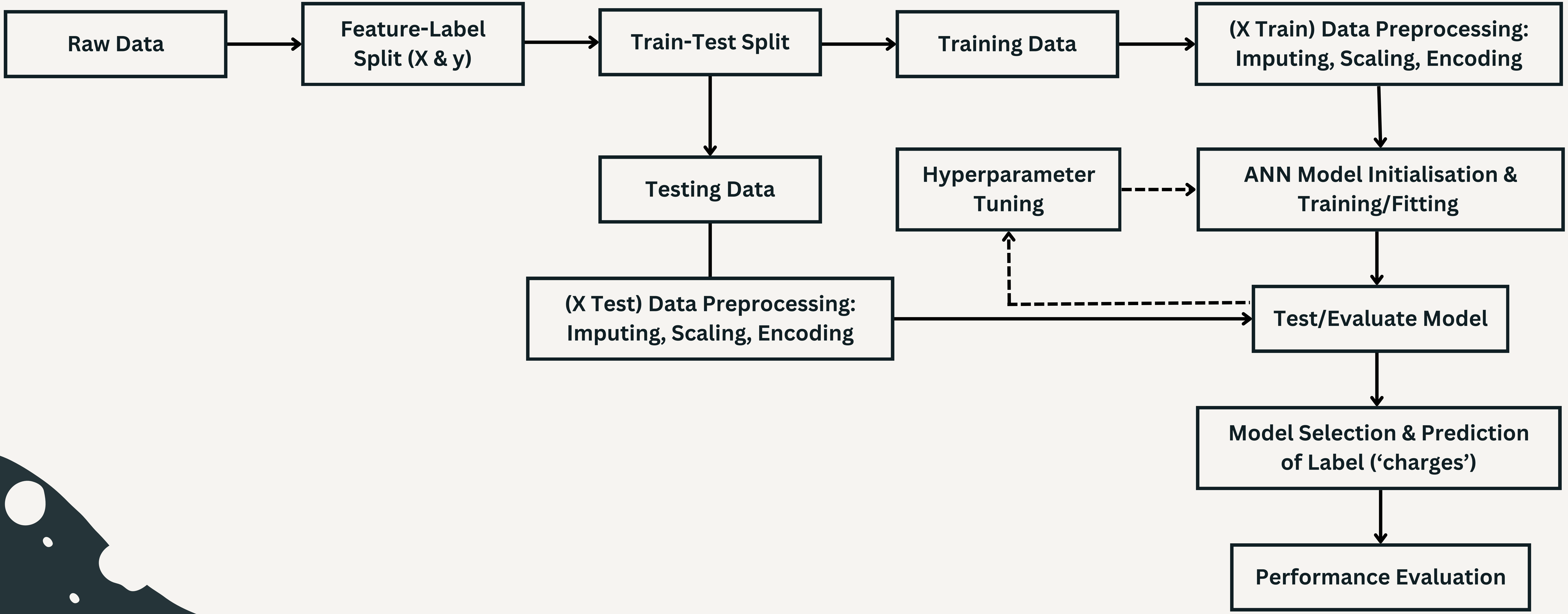
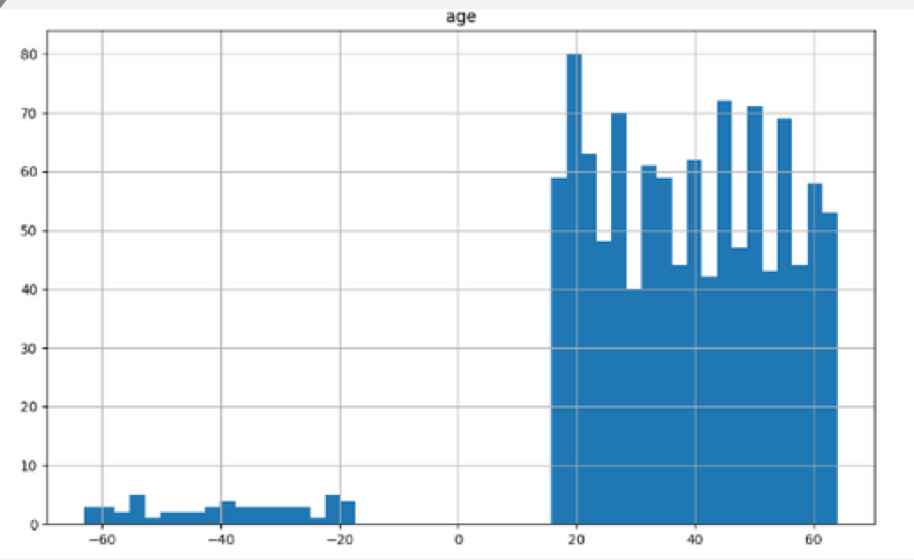
COSC202: PROJECT

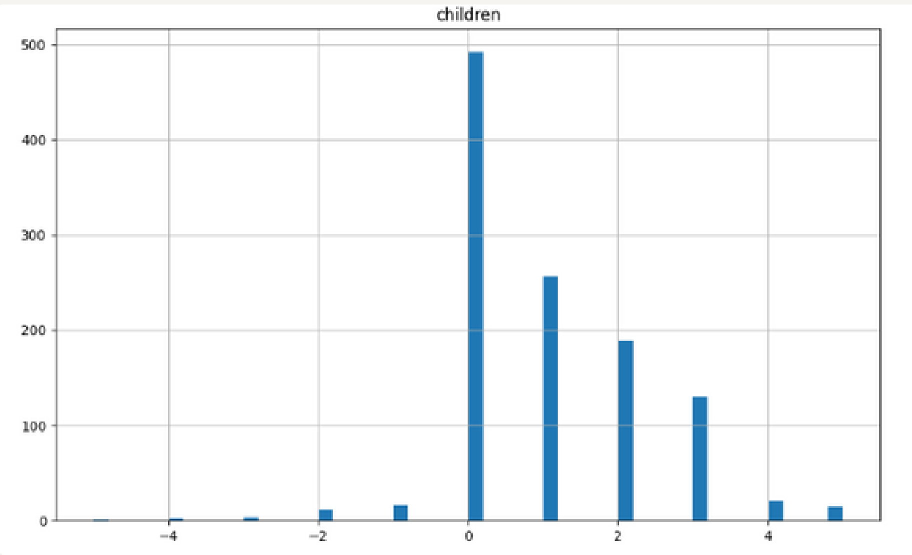
WORKFLOW DIAGRAM



DATASET VISUALIZATION

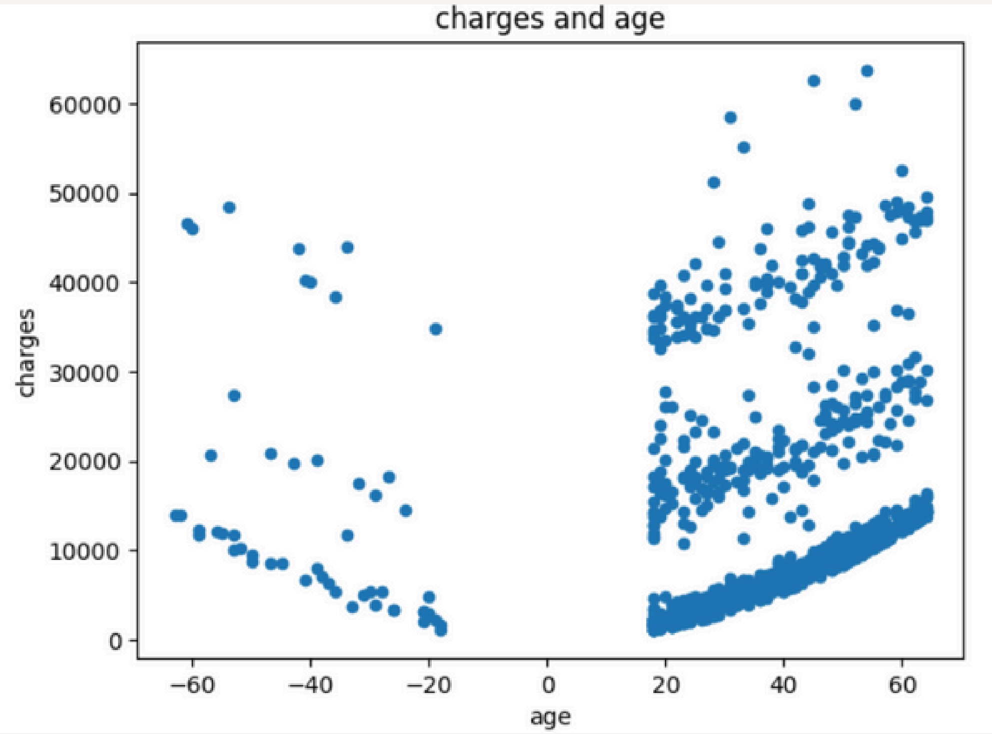
By visualising the relationships between features, our key observations include:

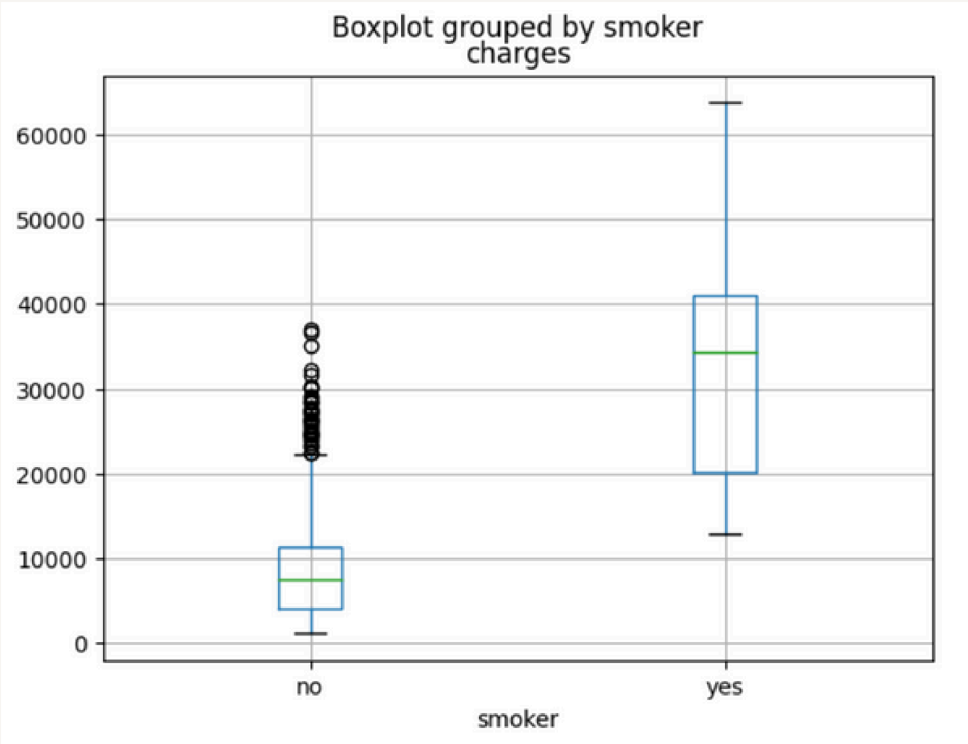


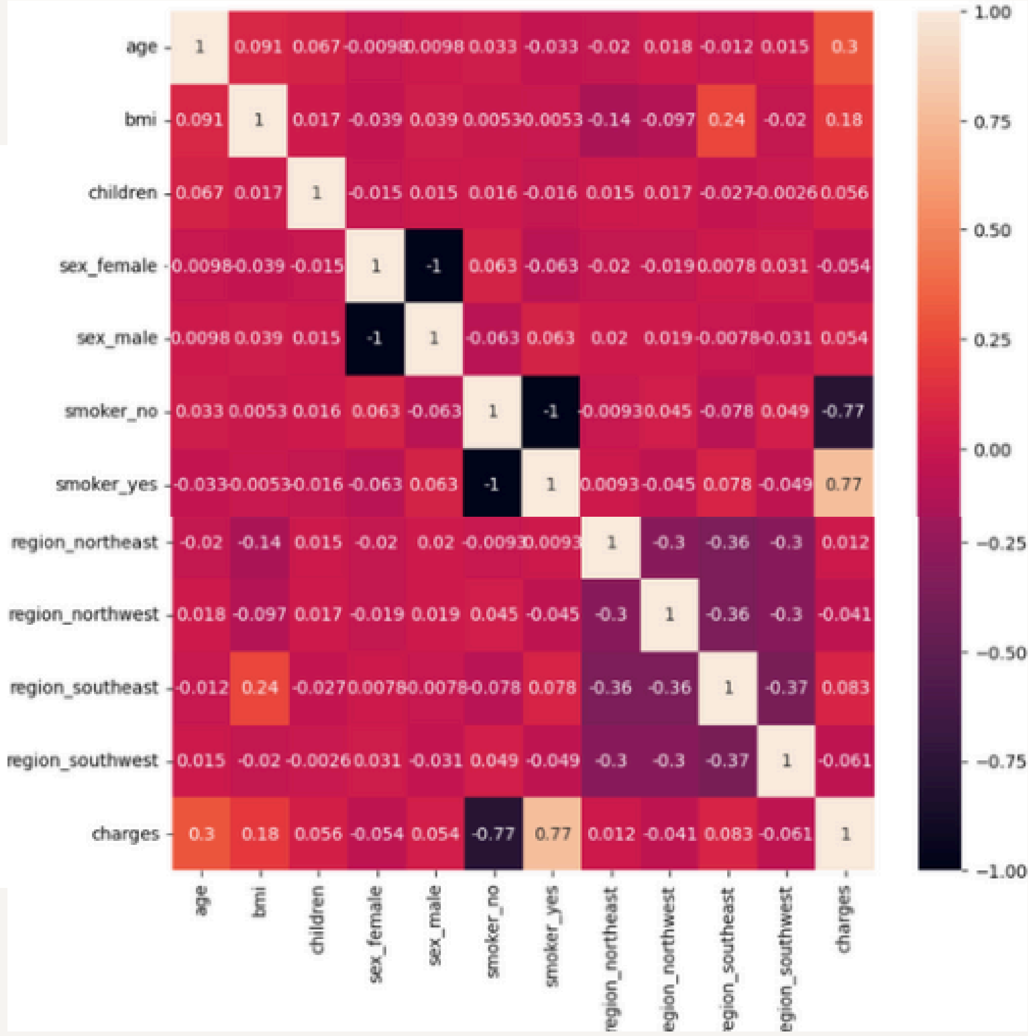


- The histograms showed **negative outliers** in **'age'** and **'children'**.

- The **scatter plot** shows **'age'** is **proportional** to **'charges',** even for the negatives.







- Both the **box plot** & **heatmap** show a **strong relationship** between **'smoking'** and **'charges'**.
- The **box plot** also shows **'age'** and **'bmi'** have **significant correlations** with **'charges'**.

PREPROCESSING

1) Outlier & Typo Handling

- Took *absolute value* of **‘age’**.
- *Mode-imputed* negative **‘children’** values.
- *Mode-imputed* repeated **‘sex’** typos.

2) Missing Data Imputation

- *Mode-imputed* missing data in **‘region’**.
- *Median-imputed* missing data in **‘bmi’**.

3) Feature Scaling

- *MinMax scaled* the numerical columns: **‘age’**, **‘bmi’**, and **‘children’**.

4) Encoding

- *One-hot encoded* the categorical columns: **‘sex’**, **‘smoker’**, and **‘region’**.

	age	bmi	children	sex_female	sex_male	smoker_no	smoker_yes	region_northeast	region_northwest	region_southeast	region_southwest
319	0.021739	0.145539	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
586	0.282609	0.590347	0.4	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1048	0.695652	0.314167	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0
246	0.086957	0.559660	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0
789	0.326087	0.557575	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...

HYPERPARAMETER TUNING

We experimented with tuning hyperparameters including the number of neurons, hidden layers, dropout rates, optimizer, and epochs.

	Model 1	Model 2	Model 3	Model 5 (Selected)
Number of Layers	Input: 1 layer with 16 neurons	Input: 1 layer with 64 neurons	Input: 1 layer with 100 neuron	Input: 1 layer with 100 neuron
	Hidden: 1 layer with 64 neurons	Hidden: 2 layers, each with 64 and 32 neurons, respectively	Hidden: 3 layers with 128, 64, and 32 neurons, respectively	Hidden: 3 layers with 128, 64, and 32 neurons, respectively
	Output: 1 neuron	Output: 1 neuron	Output: 1 neuron	Output: 1 neuron
Activation Function	ReLU	ReLU	ReLU	ReLU
Dropout Rate	10% (0.1) after each layer	10% (0.1) after each layer	10% (0.1) after each layer	10% (0.1) after each layer
Optimizer	Adam	Adam	RMSprop	RMSprop
Learning Rate	Default (0.001)	Default (0.001)	Default (0.001)	0.0005
Epochs	500	500	500	150
Batch Size	-	-	-	16

SELECTED MODEL (MODEL 5)

Chosen model ***architecture****:*

```
model5 = Sequential([
    Dense(100, activation='relu', input_shape=(11,)),  # Input layer
    Dropout(0.1),  # Regularization
    Dense(128, activation='relu'),  # Hidden layer 1
    Dropout(0.1),  # Regularization
    Dense(64, activation='relu'),  # Hidden layer 2
    Dropout(0.1),  # Regularization
    Dense(32, activation='relu'),  # Hidden layer 3
    Dropout(0.1),  # Regularization
    Dense(1)  # Single output for regression
])
```

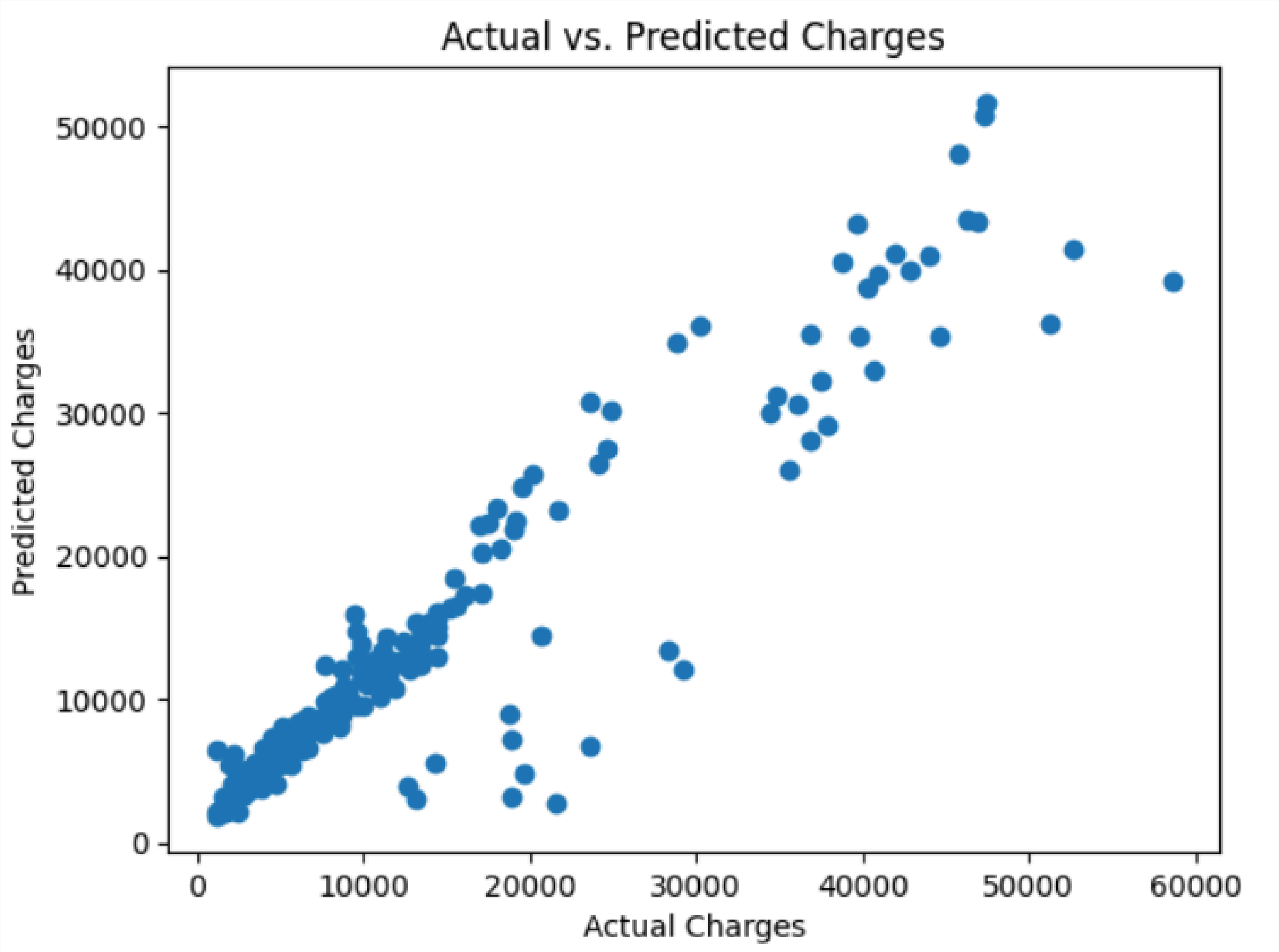
```
optimizer=RMSprop(learning_rate=0.0005)
model5.compile(
    optimizer = optimizer,
    loss='mean_squared_error',
    metrics=['mean_absolute_error']
)
history5 = model5.fit(
    X_train_transformed, y_train,
    validation_data=(X_test_transformed, y_test),
    epochs=150,
    batch_size=16
)
```

Main ***differences*** *from other models:*

- ***RMSprop*** *optimizer with a* ***learning rate*** *of 0.0005*
- ***150*** *epochs*
- ***Batch size*** *=* ***16***

MODEL PERFORMANCE

*The model has an **MAE** of **2585.68**, which is one of the **lowest** of our models. It also took the **least epochs** to get to this MAE, the others may have overfit.*



*Generally the model has **better predictions** for **lower charges**, this is due to there being **less data points with high charges** in the dataset.*

*However, the graph is **relatively linear** which means the is **performing well**.*

THANK YOU