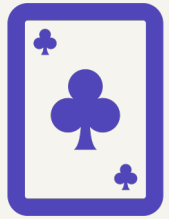
Insurance Claim Prediction

Business Challenge

This dataset contains relevant data regarding the medical cost for multiple people

The goal is to observe which variables help affect the likelihood of an insurance claim



Will then help predict the likelihood of an insurance claim for future users.

Data Pre-Processing

Remove unnecessary columns

In this case, I removed the region column as it did not pertain to my business question

Handle data anomalies

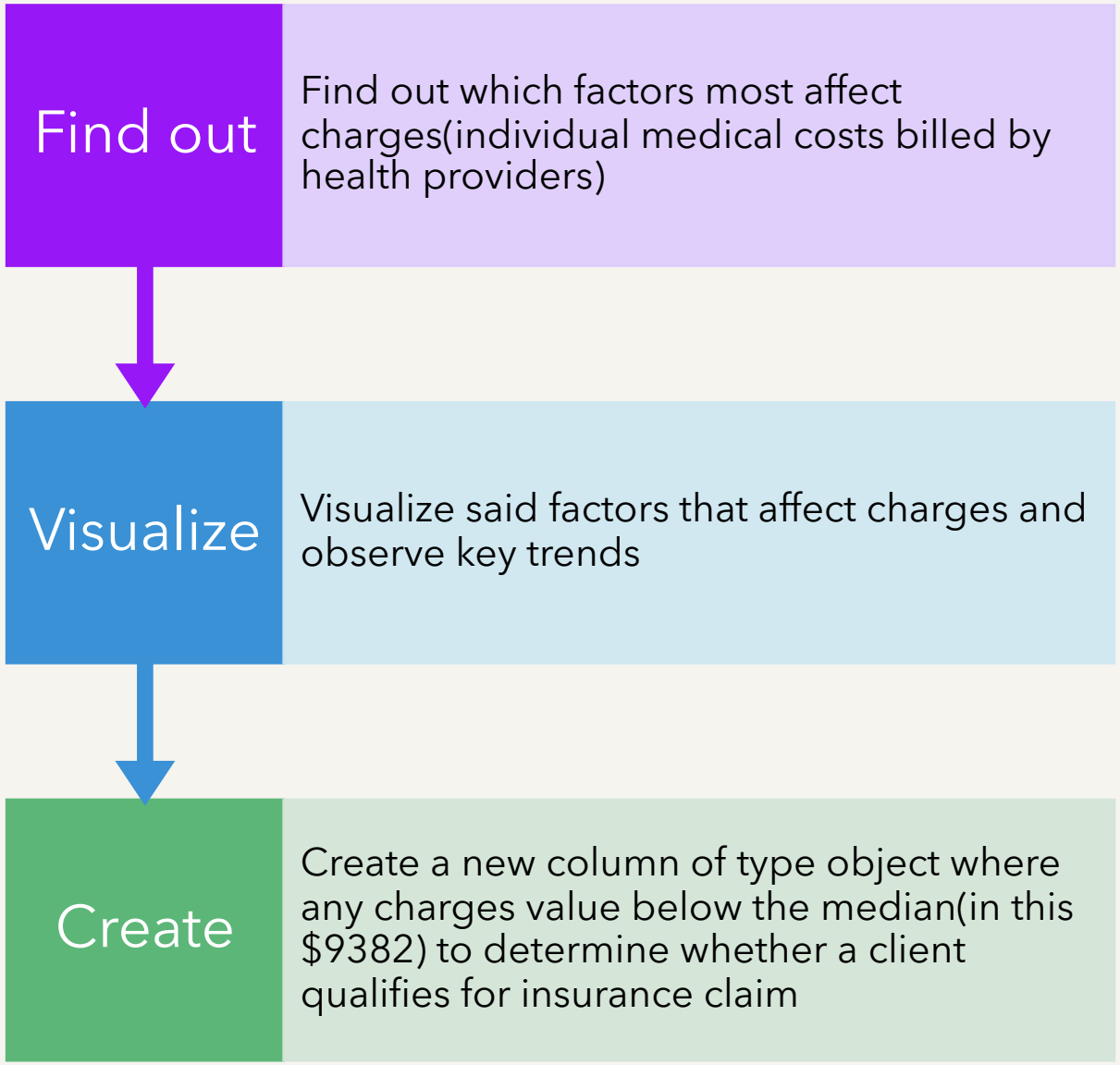
Were not any data anomalies in this dataset

```
df.isnull().sum()
```

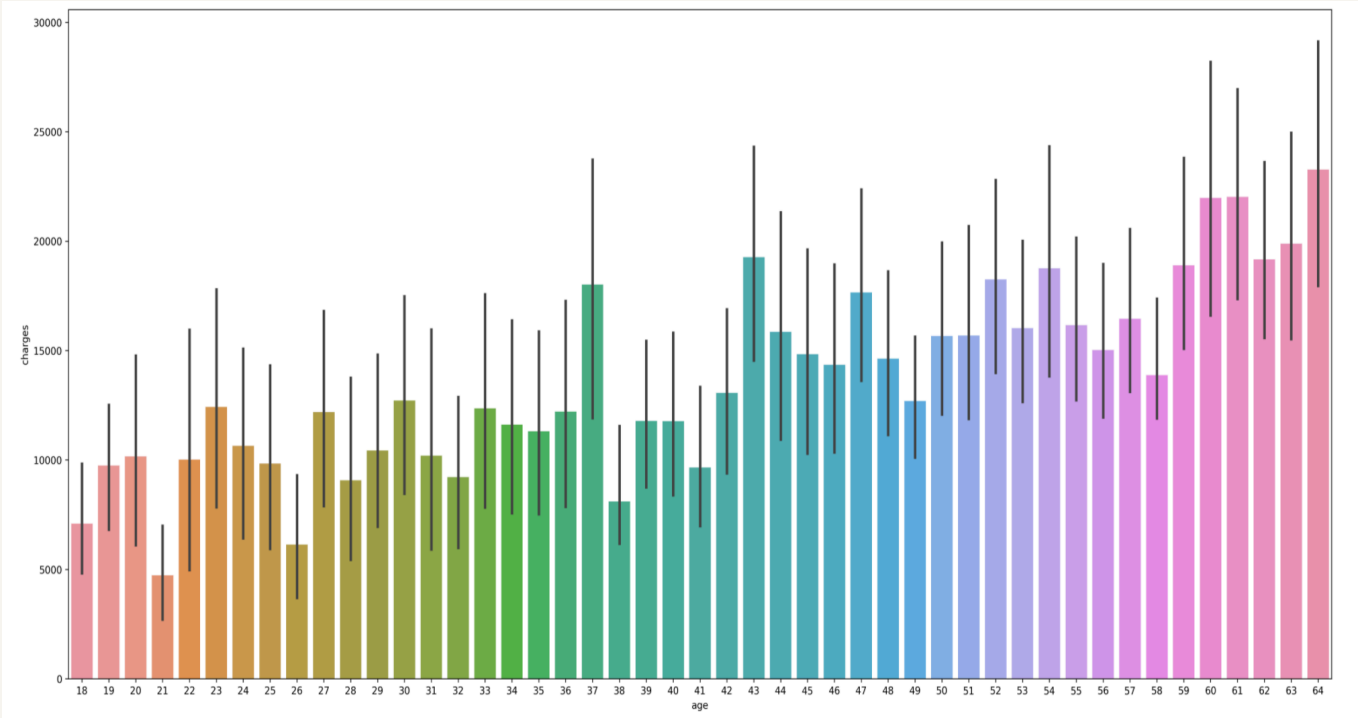
```
age         0
sex         0
bmi         0
children    0
smoker      0
charges     0
dtype: int64
```

```
In [323]: df.columns
Out[323]: Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtype='object')

In [324]: df.drop(df.columns[[5]], axis=1, inplace=True)

In [325]: df.columns
Out[325]: Index(['age', 'sex', 'bmi', 'children', 'smoker', 'charges'], dtype='object')
```

Data Exploration



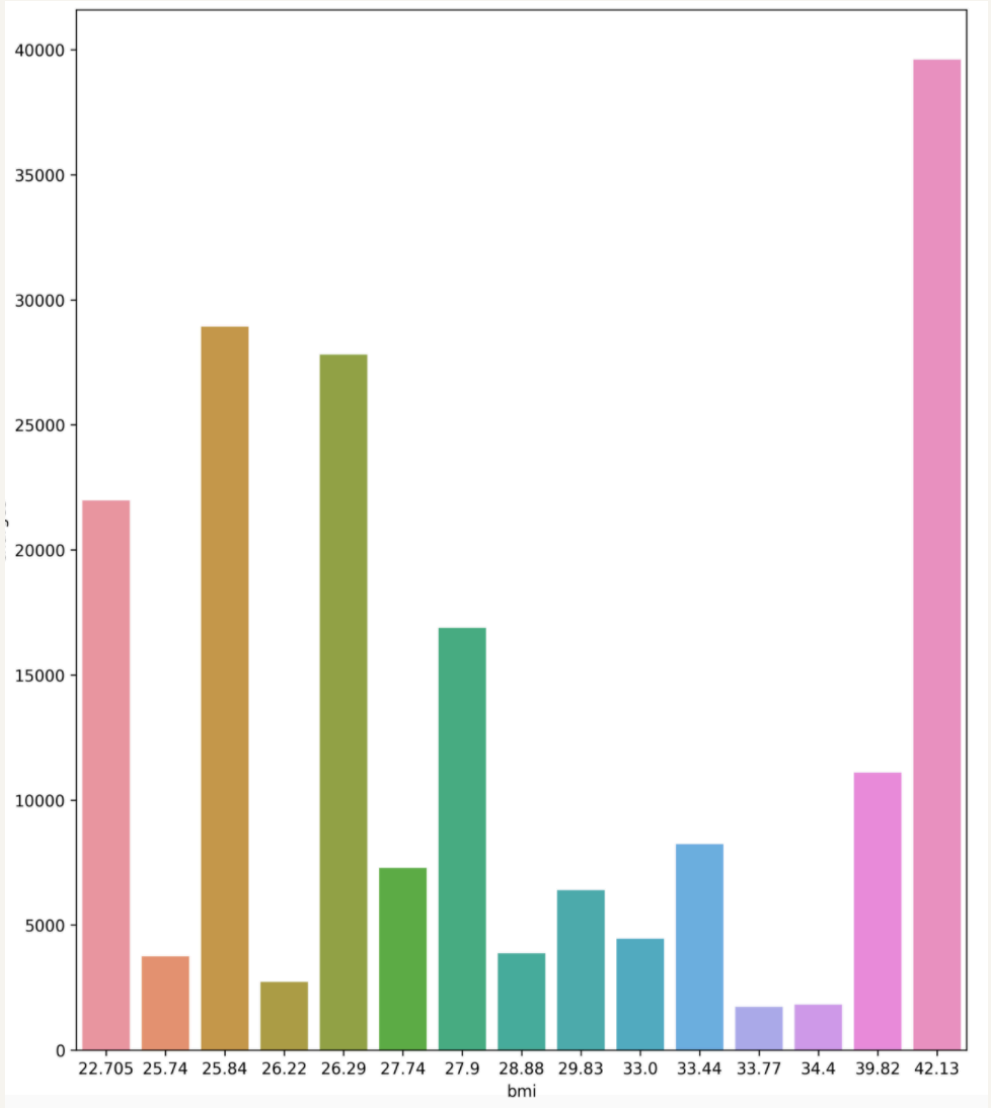
How Does Age Affect Insurance Medical Charges



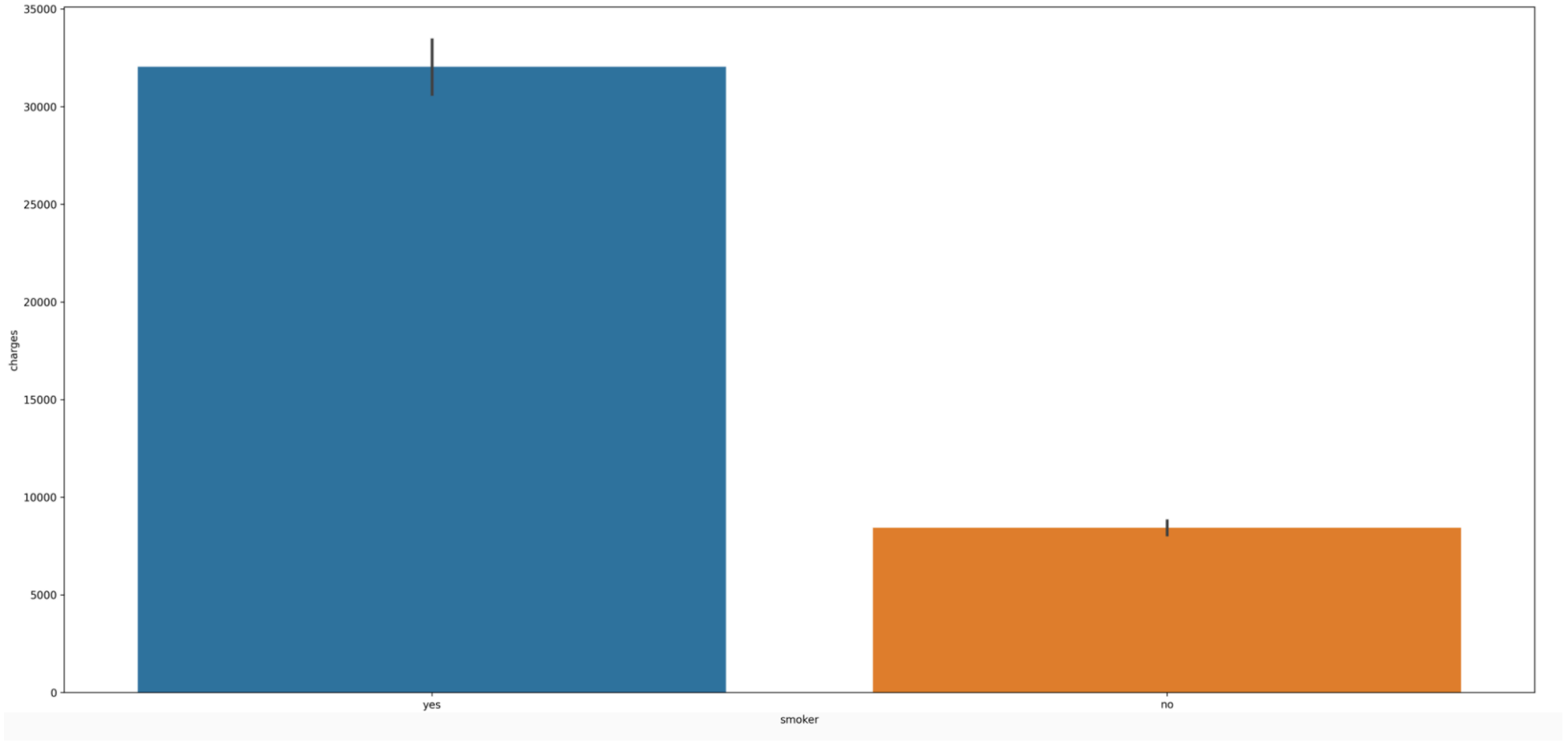
- Generally, the greater the age, the greater the insurance charge
- Checks out due to older people having more medical problems as they age

Is There A Correlation Between BMI And Insurance Charges

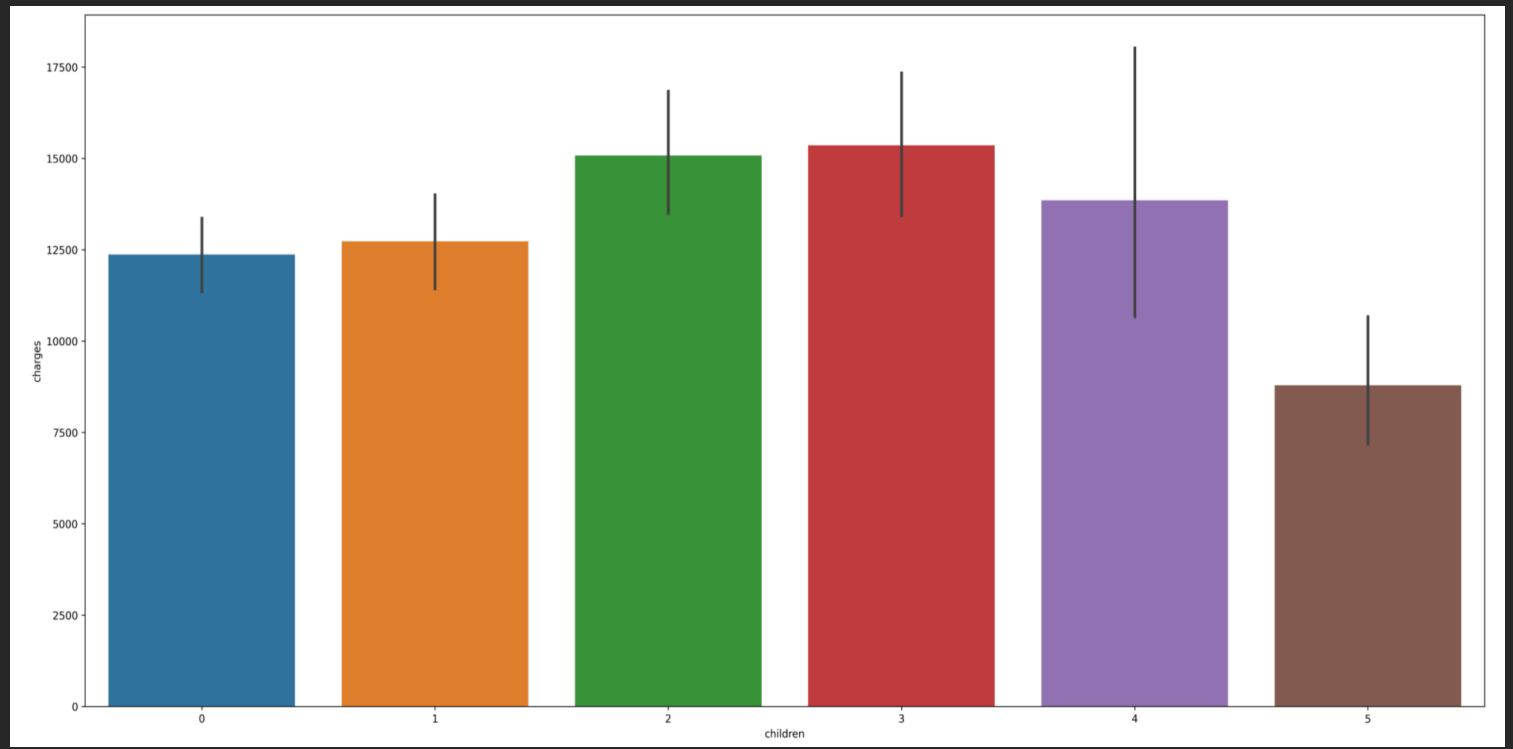
- No discernable relationship between BMI and insurance charges



Do People Who Smoke Have Higher Insurance Charges

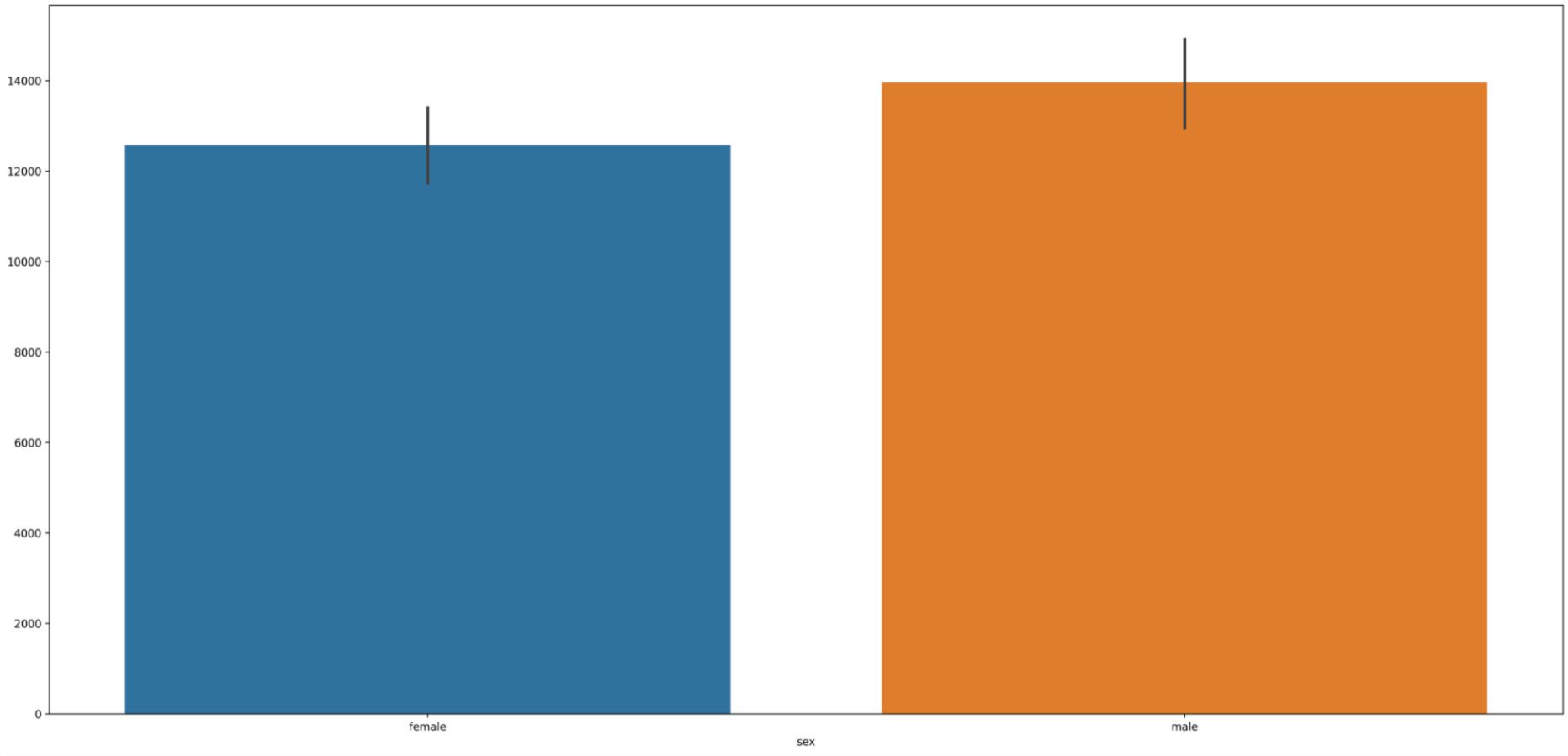


What Is The Relationship Between The Number of Children And Insurance Charges



- People with 2-3 children have the highest insurance charges
- However, dataset is slightly inconclusive due to a minuscule sample size for clients with 4 or 5 children relatively

Does A Person's Gender Affect Insurance Charges



Data Transformation

- Using various classification models predicting insurance claim based on the insurance charge amount
- Due to charges being of type float as well as the response variable, my models will not work
- Thus, I created a new column of type object and assigned a value of True(Insurance Claim) if the charge is less than $9382 and vice versa

```
df['> 9382'] = df['charges'].apply(lambda x: 'True' if x <= 9382 else 'False')

print(df)
```

	age	sex	bmi	children	smoker	charges	> 9382
0	19	female	27.900	0	yes	16884.92400	False
1	18	male	33.770	1	no	1725.55230	True
2	28	male	33.000	3	no	4449.46200	True
3	33	male	22.705	0	no	21984.47061	False
4	32	male	28.880	0	no	3866.85520	True
...	...	...	...	...	...	...	...
1333	50	male	30.970	3	no	10600.54830	False
1334	18	female	31.920	0	no	2205.98080	True
1335	18	female	36.850	0	no	1629.83350	True
1336	21	female	25.800	0	no	2007.94500	True
1337	61	female	29.070	0	yes	29141.36030	False

Data Modeling

Use machine learning techniques such as Logistic Regression, Decision Tree, and Random Forest to best help predict future insurance claims

Based on findings of EDA, my predictor variables are sex, children, smoker, children, and age and my response/target variable is the new column I created during Data Transformation

This will hopefully give my models a more accurate indication of future insurance claims

Logistic Regression

```
X_train, X_test, Y_train, Y_test = sklearn.model_selection.train_test_split(df_selected,df['> 9382'],test_size=0.30,
```

```
logreg.fit(X_train, Y_train)

Y_pred = logreg.predict(X_test)

logreg.score(X_test, Y_test)
```

```
0.8905472636815921
```

An algorithm for classification

Helps find relationships between features and a particular outcome

With a score of 0.89, my model is highly accurate, and it implies that my predictor variables will be a good indicator of future insurance claims

Decision Tree

```
print(classification_report(Y_test,predictions))
```

	precision	recall	f1-score	support
False	0.88	0.89	0.88	199
True	0.89	0.88	0.88	203
accuracy			0.88	402
macro avg	0.88	0.88	0.88	402
weighted avg	0.88	0.88	0.88	402

Supervised, non-parametric learning method that is used for classification

Helps create a model that is inferred from the data features by predicting the value of a selected variable through simple decision rules

Both recall and precision scores are significantly high. This is a very accurate model

Random Forest Classifier

- Helps select random samples from a dataset
- Using each sample, constructs a decision tree
- Then performs a vote for each sample with an associated predicted result
- Both recall and precision scores are significantly high. Highly accurate model that also means that this is an excellent indicator in predicting future insurance claims

	precision	recall	f1-score	support
False	0.93	0.89	0.91	199
True	0.90	0.94	0.92	203
accuracy			0.92	402
macro avg	0.92	0.92	0.92	402
weighted avg	0.92	0.92	0.92	402

Model Comparison And Conclusion

- Used three machine learning models: Logistic Regression, Random Forest, and Decision Trees
- Highly accurate and do lend credence to the business statement that a person's age, sex, smoking status, and number of children are great indicators of predicting future insurance claims
- Random Forest was the most accurate model with a precision and recall score of 0.93.

Any Questions?