End to End Data Science

Medical Cost Analysis - Insight and Prediction

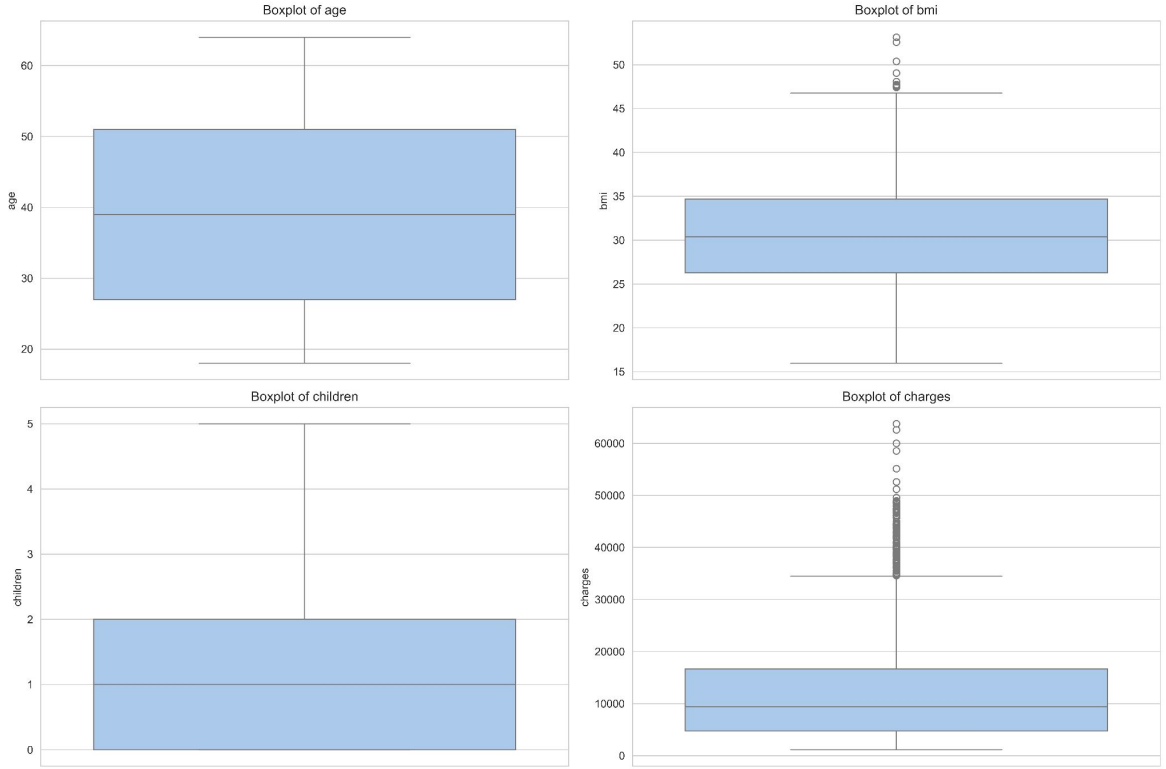
Zhoubin Maneshi
18.04.2025

Overview

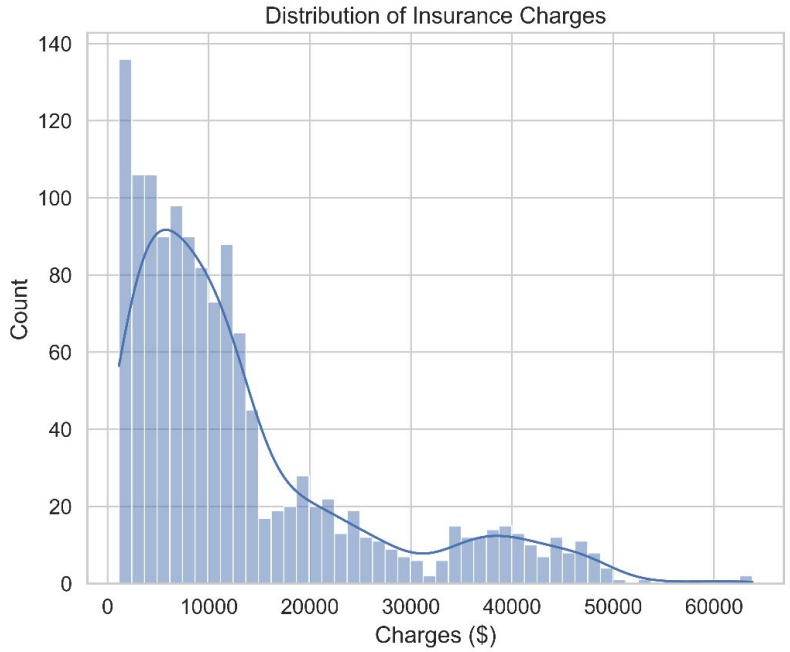
- **Dataset**
  - **Source:** https://www.kaggle.com/datasets/mirichoi0218/insurance
  - **1.3k+** rows, **7** columns
  - **Clean** by default (no NaNs, no missing vals, no special chars)
- **Features**
  - **Categorical:** sex: [‘M’, ‘F’], smoker: [‘Y’, ‘N’], region: [‘SW’, ‘SE’, ‘NW’, NE’]
  - **Numerical:** age, bmi, children
- **Supervised ML (Regression)**
  - **Target:** charges (continuous numeric)

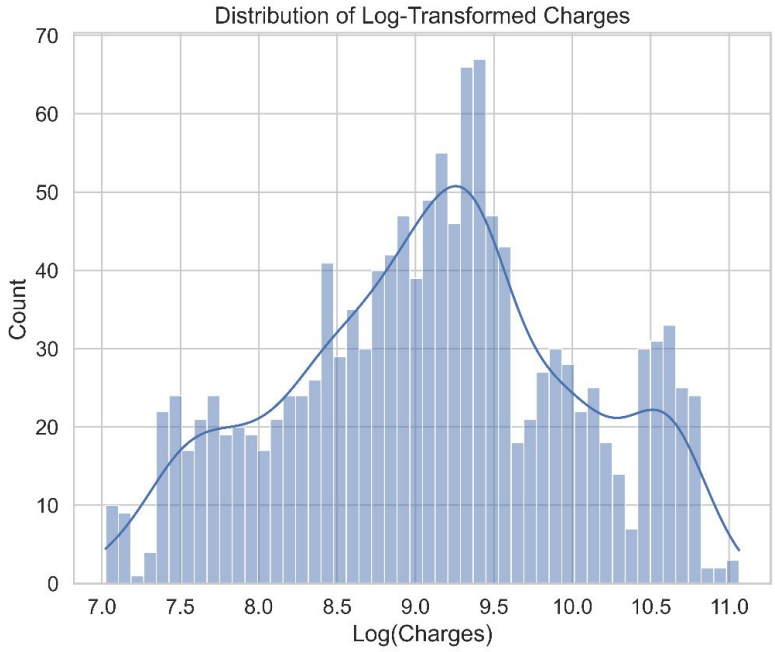
Data Exploration & Inferential Statistics

EDA: Boxplot of Numeric Values

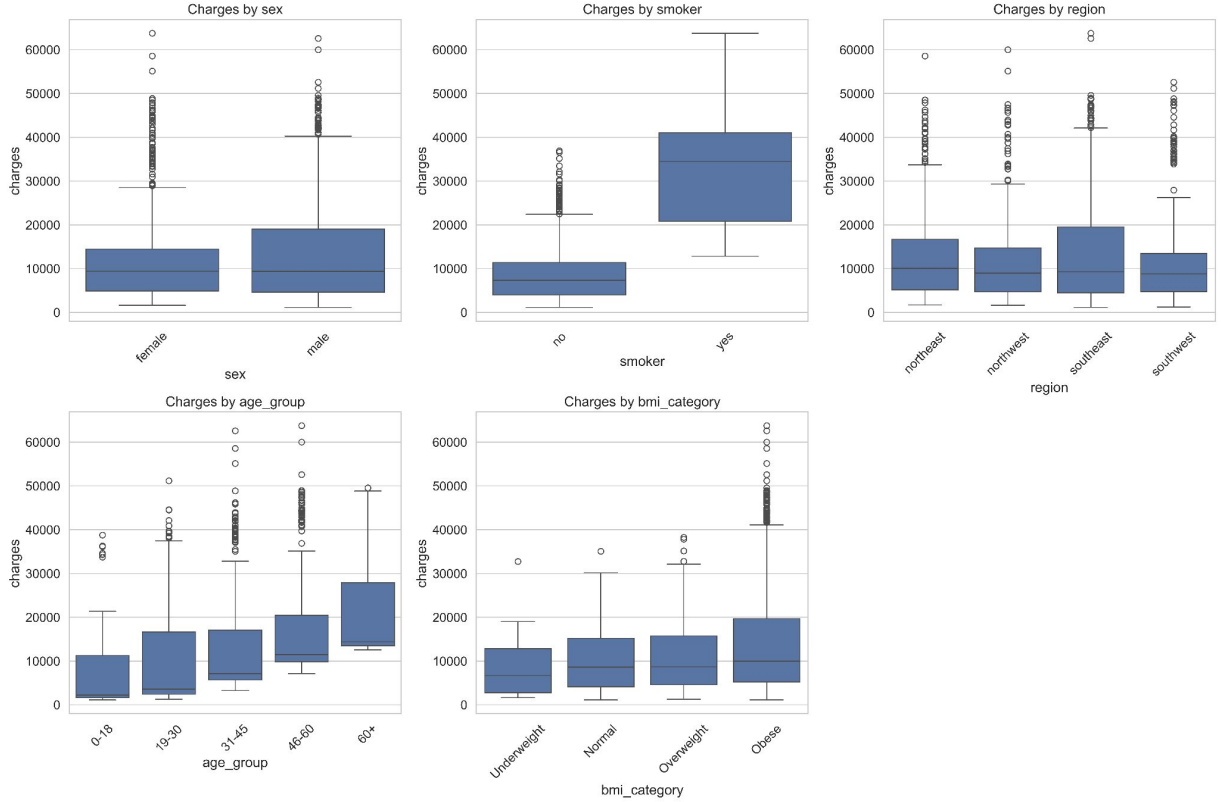


EDA: Distribution of Charges

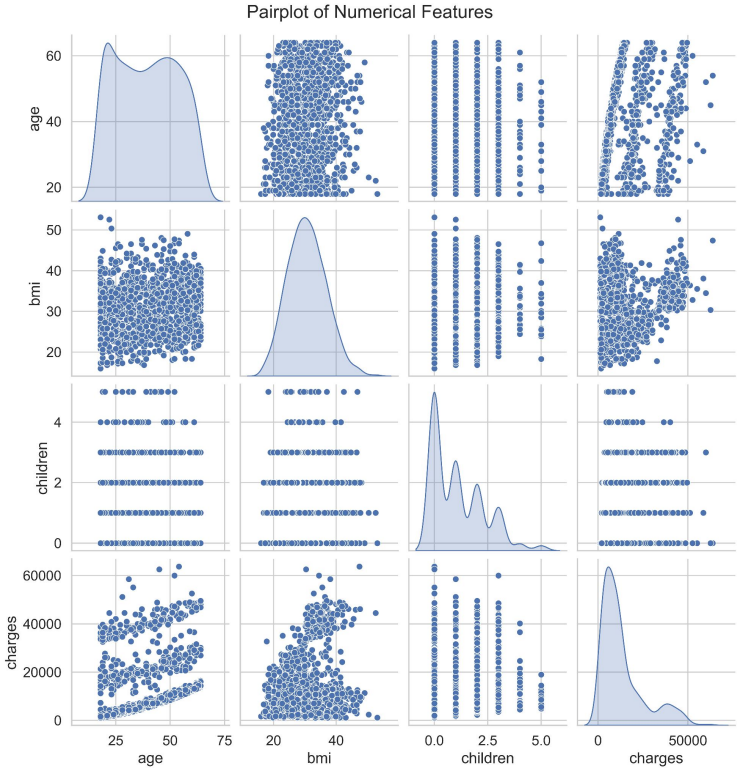


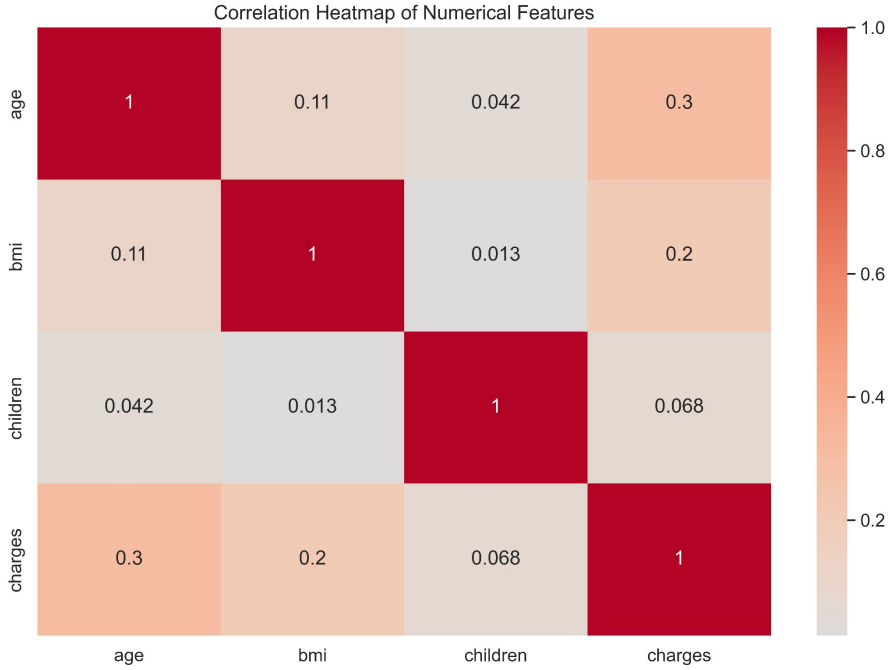


EDA: Charges by Categorical Features



EDA: Pairplot & Heatmap of Numerical Features





Inferential Statistics

- **T-Test Analysis for Charge Differences btw Smokers & Non-smokers**
  - **Hypothesis:** Smokers have significantly higher charges
  - **T-statistic:** 32.75, **p-value:** 0.0000
  - **Conclusion:** hypothesis validated ($23,615.96 difference)
- **Anova Test for Charges Across BMI Categories**
  - **Hypothesis:** There are significant differences in charges across BMI categories
  - **F-statistic:** 18.80, **p-value:** 0.0000
  - **Conclusion:** hypothesis validated

Inferential Statistics

- **Pearson Correlation btw Age and Charges**
  - **ρ:** 0.30
  - **p-value:** 0.0000
- **Chi-square Test for Smoking and Sex**
  - **Chi2-statistic:** 7.39**, p-value:** 0.0065
  - **Conclusion:** There is a significant association btw sex and smoking status

Preprocessing & Feature Engineering

Preprocessing Summary

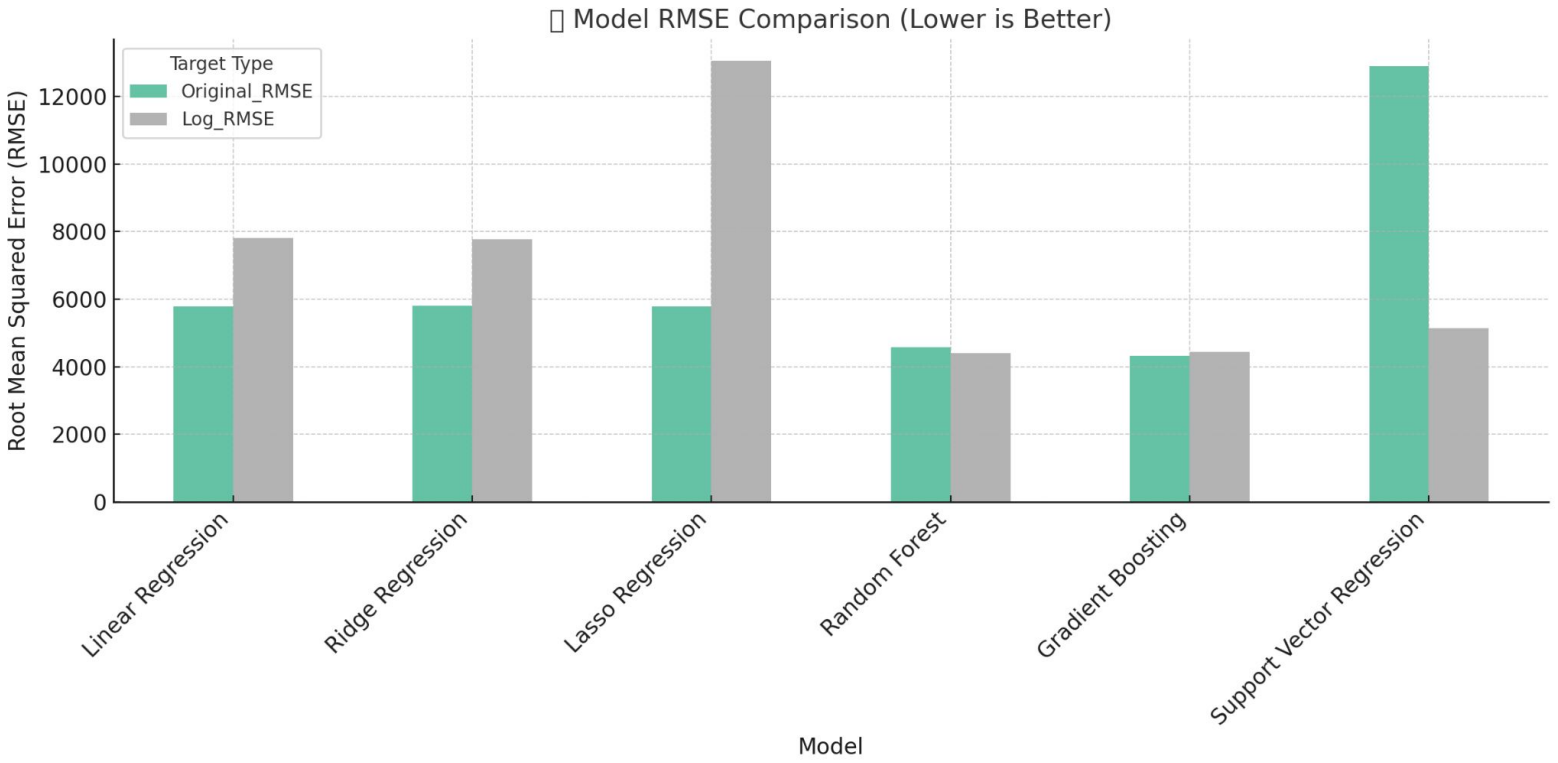
- **Target**
  - **charges** (original) and **log_charges** to handle skewness
- **Selected Features**
  - age, sex, bmi, children, smoker, region
- **Train-Test Split**
  - **80% training & 20% testing**, done separately for both original and log-transformed
- **Preprocessing Pipeline**
  - **Standardized** numerical and **One-hot encoded** categorical
- **Pipeline Assembly**
  - **ColumnTransformer** to combine preprocessed numeric and categorical
- **Feature Selection**
  - **SelectKBest** used with **f-regression** to evaluate all features for their relevance to target

Model Training & Evaluation Summary

Performance Summary

		RMSE	MAE	R2
Gradient Boosting	Original Target	4328.147789	2404.901760	0.879336
	Log Target	4447.935092	2057.149873	0.872565
Random Forest	Original Target	4582.972573	2541.614594	0.864710
	Log Target	4399.946564	2080.646795	0.875300
Linear Regression	Original Target	5796.284659	4181.194474	0.783593
	Log Target	7814.064026	3888.443159	0.606698
Lasso Regression	Original Target	5797.054261	4182.081076	0.783536
	Log Target	13053.719350	8603.157345	-0.097591
Ridge Regression	Original Target	5798.298795	4186.913072	0.783443
	Log Target	7780.621059	3881.879523	0.610058
Support Vector Regression	Original Target	12892.023995	8605.845654	-0.070568
	Log Target	5140.656196	2302.168857	0.829781

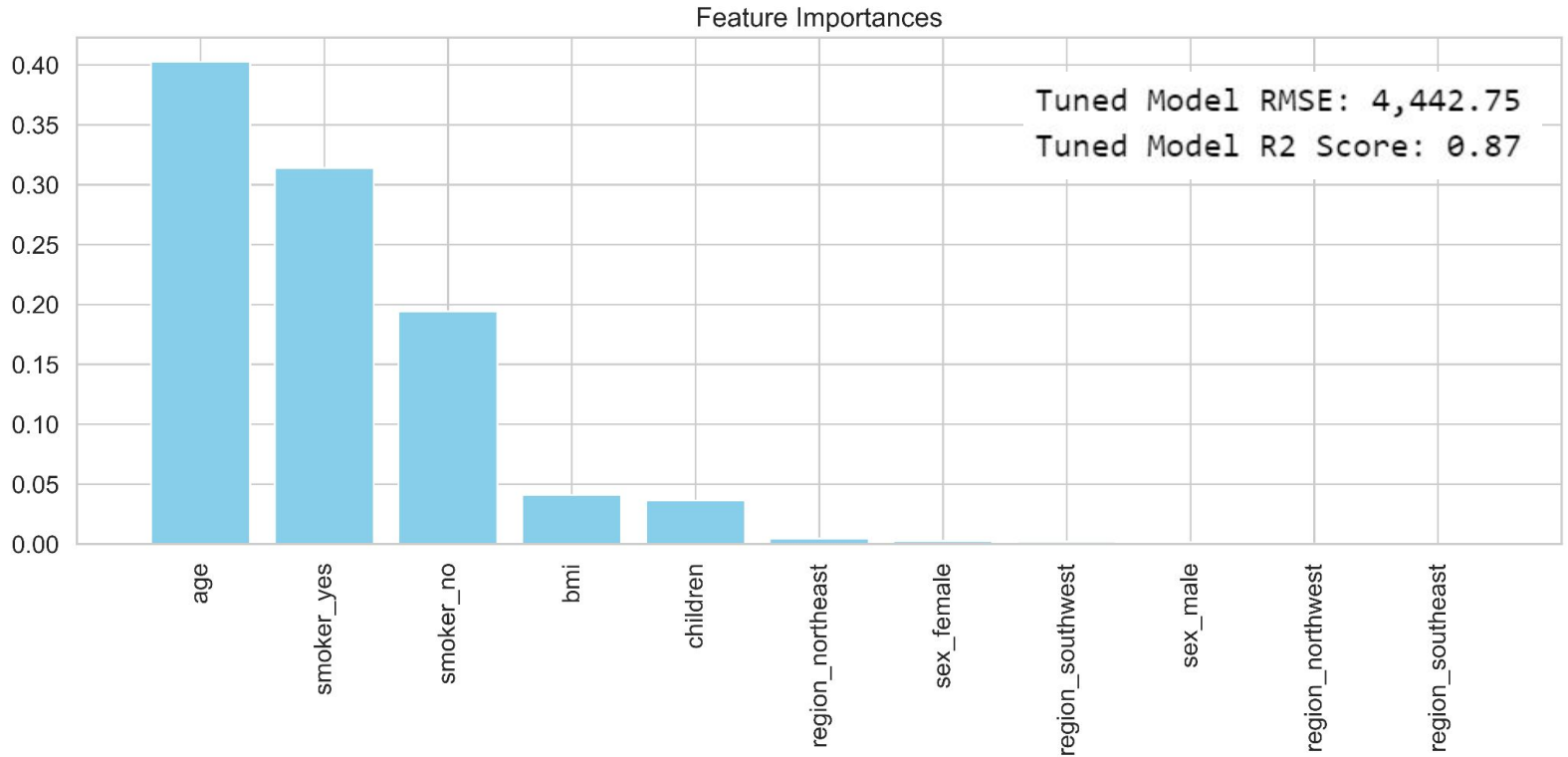
Performance Comparison Chart



Hyperparameter Tuning

- **Objective**
  - Fine-tune the best model (GBR w/ Log-Tr.) to improve predictive power
- **Steps**
  - Parameter Grid Defined (trees no., step size shrinkage, depth, node split)
  - GridSearchCV Applied (5-fold cross validation)
  - Best Parameters Identified (optimal settings from the grid search)
  - Used test to compute RMSE and R^2 score
  - Feature Importance Visualization (extracted feature names and ranked by importance)

Feature Importance



Conclusion & Key Takeaways

- **Medical Insights**
  - **Age** & **Smoking Status** are the most impactful predictors of medical costs
  - Strong statistical evidence supports significant cost variations across lifestyle and demographic factors
- **Modeling Outcome**
  - **GBR (Log Charges)** achieved the best performance
  - Hyperparameter tuning and feature engineering improved predictive accuracy
- **Next Steps**
  - Add **more features** for better performance scores
  - Extend to **time-series** or **cost forecasting** over years
  - Incorporate **external health data** for richer insights

Thank you!

Questions?