**Medical Costs Analysis — Exploratory Data Analysis, Preprocessing, and Insights**

**Dataset:** Medical Cost Personal Datasets (insurance.csv)

**Planned idea and goal:**

The main idea of this project is analyzing medical insurance costs based on various factors such as age, gender, BMI, number of dependents, smoking habits, and residential region. The goal is to explore the “Medical Cost Personal Datasets” to uncover patterns and insights that could inform pricing strategies, risk assessment, and personalized health care planning.

**Executive Summary**

This report summarizes an exploratory data analysis (EDA) and preprocessing pipeline performed on the Medical Cost Personal dataset. The objective was to understand how demographic and health-related features (age, sex, BMI, children, smoking status, region) relate to annual insurance charges, to prepare the data for modeling, and to surface actionable insights for stakeholders.

Key findings (high level):

* **Smoking status** is the strongest predictor of charges — smokers pay substantially higher average charges than non-smokers.
* BMI correlates with charges, particularly among smokers: higher BMI + smoker → much higher costs. BMI (Body Mass Index) is a simple calculation of a person's **weight** in relation to their **height** to estimate body fat, serving as a screening tool for potential health risks associated with being underweight, overweight, or obese.
* **Age** shows moderate positive correlation with charges (older individuals tend to have higher charges), but its effect is smaller than smoker status and some regional patterns.
* **Region** shows small-to-moderate differences in mean charges; differences may reflect demographic or healthcare-cost differences in regions.

**1. Data Overview**

* **Rows:** 1,338
* **Columns:** 7 — age, sex, bmi, children, smoker, region, charges
* **Types:** numeric (age, bmi, children, charges), categorical (sex, smoker, region)
* **Missing values:** none detected (0 missing entries in all columns)

The dataset is clean in terms of completeness but contains categorical text columns that require encoding before modeling.

**2. Exploratory Data Analysis (EDA)**

**2.1 Descriptive statistics (numerical)**

* age: distribution across adult ages, mean and spread showing typical insurance-age distribution.
* bmi: distribution with many observations in the overweight range; presence of higher BMI outliers.
* charges: strongly right-skewed with a small set of high-cost outliers (large medical expenses).

**2.2 Categorical summaries**

* sex: roughly balanced between male and female.
* smoker: minority are smokers, but they account for a disproportionate share of high charges.
* region: four categorical regions with similar row counts (northeast, northwest, southeast, southwest).

**2.3 Visualizations produced**

(Displayed in the analysis; recommended to incorporate into the presentation)

* Histograms: age, bmi, charges to show distributions and skewness.
* Boxplots: charges by smoker (clear separation), sex, and region (some variation).
* Scatter plots: bmi vs charges colored by smoker status — showed notable difference in slopes.
* Correlation heatmap: after mapping categorical variables to numeric, correlation of charges with features was estimated (smoker, bmi, age, children, region indicators).

**2.4 Key EDA insights**

* **Smoking is the most dominant feature**: median and mean charges for smokers are substantially higher than for non-smokers; smokers are frequent among top charge outliers.
* **BMI matters more for smokers**: scatter plots show that as BMI increases, smokers' charges increase steeply, while non-smokers show a milder increase.
* **Age has a positive relationship** with charges, but less pronounced than smoking and BMI.
* **Children**: little direct visual correlation with charges; small positive or neutral effect.
* **Sex**: small differences after controlling other variables; not a primary driver.

**3. Data Preprocessing Steps**

These steps were implemented in the pipeline and saved to insurance\_processed.csv.

**3.1 Missing values**

* Checked for missing values. None were present.
* If missing values existed, plan: numeric → median imputation; categorical → mode imputation.

**3.2 Encoding categorical variables**

* sex mapped to binary: male → 1, female → 0.
* smoker mapped to binary: yes → 1, no → 0.
* region converted via **one-hot encoding** (four columns region\_northeast, region\_northwest, region\_southeast, region\_southwest).

Rationale: many models require numeric inputs; one-hot preserves region-specific effects while avoiding implicit ordinal assumptions.

**3.3 Feature scaling**

* Standardized numeric features age, bmi, and charges using StandardScaler (result: mean 0, std 1 for each scaled column).
* Rationale: scaling improves training stability and removes scale-driven bias for models sensitive to feature scales (e.g., regularized linear models, KNN, SVM, neural nets).

**3.4 Output**

* The processed dataframe contains encoded region columns, numeric-coded binary features, and standardized numeric columns.

**4. Feature Exploration**

**4.1 Correlation analysis**

* After encoding, smoker had the highest positive correlation with charges.
* bmi and age had positive correlations too; regional dummies had mild correlations.

**5. Regional Analysis (Bonus)**

* Computed mean, median, and count of charges by region.
* Observed moderate regional variation in mean charges. Differences could reflect: regional pricing differences, demographic composition (age/BMI/smoker concentration), or sample variability.

**6. Recommendations & Next Steps**

**Modeling:**

1. Train and compare models: Random Forest, XGBoost, and regularized linear models (Ridge/Lasso). Evaluate with RMSE, MAE, and R² via cross-validation.
2. Include interaction features (smoker\*bmi, age\*bmi, region\*smoker). Test polynomial features for bmi and age.
3. Investigate log-transforming charges to stabilize skewness and improve model residuals.

**Diagnostics & fairness checks:**

* Check residuals for heteroscedasticity and non-normality.
* Validate models across subgroups (region, sex, smoker status) to ensure consistent performance and to detect biases.

**Presentation & dashboard:**

* Prepare slides summarizing EDA visuals, top features, model performance, and recommended actions.
* Optional interactive dashboard (Tableau / PowerBI / Dash / Streamlit) using insurance\_processed.csv: allow stakeholders to filter by region, smoker status, BMI, and view changes in mean predicted charges.

**Business actions:**

* Because smoking is a major cost driver, consider targeted smoking cessation programs and prevention initiatives as a cost-control lever.
* For high BMI populations, design preventive health initiatives; evaluate the ROI of lifestyle interventions.

**7. Appendix A — Code snippets (key steps)**

**Loading**

import pandas as pd

df = pd.read\_csv('/mnt/data/insurance.csv')

**One-hot encoding region**

df = pd.get\_dummies(df, columns=['region'], drop\_first=False)

**Binary mapping for sex and smoker**

df['sex'] = df['sex'].map({'male': 1, 'female': 0})

df['smoker'] = df['smoker'].map({'yes': 1, 'no': 0})

**Standardization**

from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

df[['age','bmi','charges']] = scaler.fit\_transform(df[['age','bmi','charges']])

**Quick linear regression**

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(X\_train, y\_train)

coeffs = pd.Series(model.coef\_, index=X.columns).sort\_values(key=lambda s: np.abs(s), ascending=False)

**Presentation: Slide Deck Outline (for stakeholders)**

Below is a suggested 12-slide deck with speaker notes. You can copy the visuals from the EDA script into the slides.

**Slide 1 — Title**

* Title: Medical Costs Analysis — EDA & Key Insights
* Subtitle: Dataset: Medical Cost Personal Datasets
* Author, date

**Slide 2 — Executive Summary (1 slide)**

* 3–5 bullet summary of findings (smoking biggest driver, BMI interactions, regional differences, R² ~0.78)

**Slide 3 — Dataset Overview**

* Rows, columns, features, no missing values
* Short note on preprocessing steps performed

**Slide 4 — Distributions: Age, BMI, Charges**

* Histograms laid out side-by-side
* Note skewness and outliers

**Slide 5 — Smoking vs Non-smoking (Boxplots)**

* Boxplot of charges by smoker status
* Speaker note: smokers have much higher median and more extreme outliers

**Slide 6 — BMI vs Charges by Smoker Status**

* Scatter plot with smoker color-coded
* Speaker note: clear amplification effect for smokers

**Slide 7 — Correlation Matrix & Feature Importance**

* Heatmap + top linear regression coefficients
* Speaker note: smoker highest coefficient, then age, bmi

**Slide 8 — Regional Analysis**

* Table or bar chart: mean charges by region
* Discuss potential causes and need for stratified analysis

**Slide 9 — Modeling Results (Quick baseline)**

* Linear regression R² = 0.78
* Mention candidate models to try next (RF, XGBoost, regularized linear)

**Slide 10 — Recommendations**

* Short- and medium-term actions: prevention programs, modeling improvements

**Slide 11 — Dashboard Demo & Next Steps**

* Screenshot or mockup if available; features to include (filters, predicted charge viewer)

**Slide 12 — Appendix / Questions**

* Link to processed CSV and code
* Contact info

**Optional: Interactive Dashboard Plan**

**Platform choices**

* Tableau / PowerBI: best for business stakeholders; quick to build with processed CSV.
* Streamlit / Dash: easier for light web deployment and deeper custom ML integration.

**Suggested components**

* Filters: region, smoker status, age range, BMI range
* Visuals: histogram of charges, boxplot by group, scat

**Medical Costs Analysis — Business Impact Report**

**Problem Definition and Business Relevance**

Healthcare providers and insurers face significant challenges in forecasting **medical costs** and managing **risk pools**. Rising costs, particularly driven by lifestyle factors such as smoking and obesity, place pressure on both companies and customers. Without accurate insights, insurers risk mispricing premiums, underestimating high-risk groups, or failing to identify opportunities for preventive healthcare interventions.

The problem addressed in this project: **How do demographic and lifestyle features (age, sex, BMI, smoking status, dependents, region) influence medical costs, and how can insurers use these insights to improve decision-making?**

**Our Solution and Technology**

We analyzed the **Medical Cost Personal Dataset** using advanced data analytics and preprocessing techniques. The solution leverages **Python (Pandas, NumPy, Seaborn, Matplotlib)** to:

1. **Explore the dataset (EDA):** Identified cost drivers and patterns.
2. **Preprocess the data:** Encoded categorical features (sex, smoker, region), standardized continuous variables (age, BMI, charges).
3. **Visualization:** Created clear dashboards and charts to communicate insights (e.g., smokers’ charges vs non-smokers, BMI impact by smoker status).

**Key Features & Insights**

* **Smoking status**: The single most important predictor; smokers’ costs are several multiples higher than non-smokers.
* **BMI (Body Mass Index):** Strong influence on charges, especially when combined with smoking.
* **Age:** Predictable upward trend with higher costs in older populations.
* **Children:** Little direct effect, but useful for family coverage segmentation.
* **Region:** Minor but detectable variation; indicates regional pricing and demographic differences.

**Challenges and Lessons Learned**

* **Data skewness:** Charges are heavily skewed due to extreme high-cost patients. Solution: considered log-transformation to stabilize variance.
* **Interactions:** Simple models underestimate non-linear effects (e.g., smoker \* BMI). Future models need interaction terms or tree-based methods.
* **Interpretability vs accuracy tradeoff:** Linear regression is easy to explain but leaves variance unexplained; more complex models could improve predictions.

Key lesson: **Simple features (smoking, BMI, age) can already explain most cost variability.** However, **interactions and advanced models** are necessary for accurate premium forecasting.

**Potential Business Value**

**For Insurers:**

* **Risk-adjusted pricing:** Incorporate smoking and BMI into premium models for fairer, more sustainable pricing.
* **Preventive health programs:** Data shows smoking cessation and weight management programs can significantly reduce long-term claims.
* **Regional strategies:** Regional differences in mean charges highlight opportunities for tailored marketing and cost management strategies.

**For Healthcare Providers:**

* Prioritize interventions in **high-BMI smokers** — group with disproportionate costs.
* Align wellness programs with cost drivers to maximize ROI.

**Pitch Tips & Demonstration Plan**

1. **Start with the problem:** Rising medical costs threaten profitability and affordability.
2. **Show the solution:** Use data-driven insights and predictive modeling to anticipate charges.
3. **Highlight results:** Smoking = top driver, BMI and age significant, regional variations exist.
4. **Demo (if live):** Show interactive dashboard — filter by smoker status and region to instantly see impact on charges.
5. **End with impact:** Smarter premium models, better preventive programs, healthier customers, and stronger profitability.

**Conclusion**

Our project demonstrates that **data-driven analysis of medical costs** is not only feasible but highly valuable for insurers and healthcare providers. By applying EDA, preprocessing, and modeling, we have identified key cost drivers, challenges, and actionable insights that can guide **business strategy, pricing, and prevention programs**.

**Next steps:** Extend modeling to advanced methods (Random Forest, XGBoost), deploy insights in an interactive dashboard, and partner with business teams to integrate these findings into operational decision-making.