Data Busters

Charles, Clarence, Daeren, Jay

NUS SDS MINI HACKATHON 2025

Problem Statement

Healthcare costs are rising, but many are puzzled why insurance costs vary so widely.

Two people from the U.S. can pay drastically different rates - what’s driving the difference?

CHALLENGE

CAN WE USE DEMOGRAPHIC AND LIFESTYLE DATA TO PREDICT MEDICAL INSURANCE CHARGES AND UNCOVER THE KEY FACTORS THAT INFLUENCE COST?

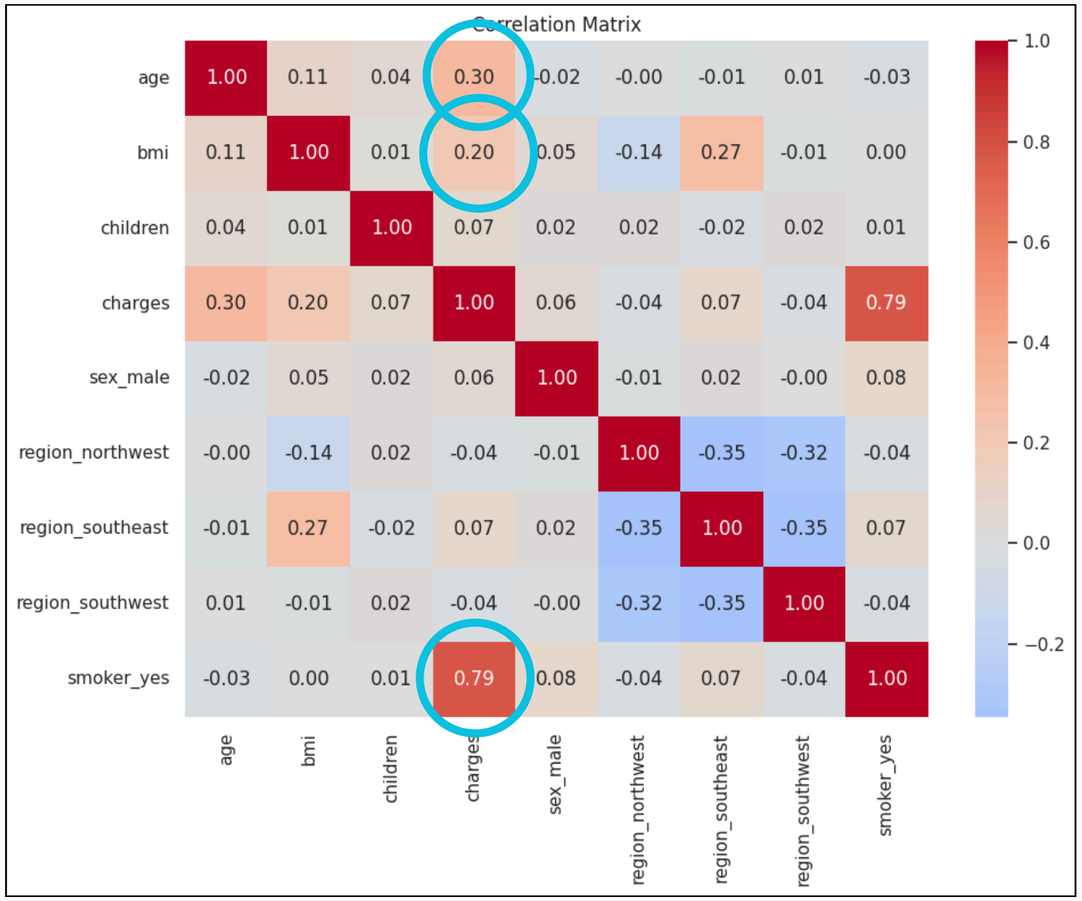
OBJECTIVES

Build a predictive model for accurate insurance cost estimation

Identify the most influential variables

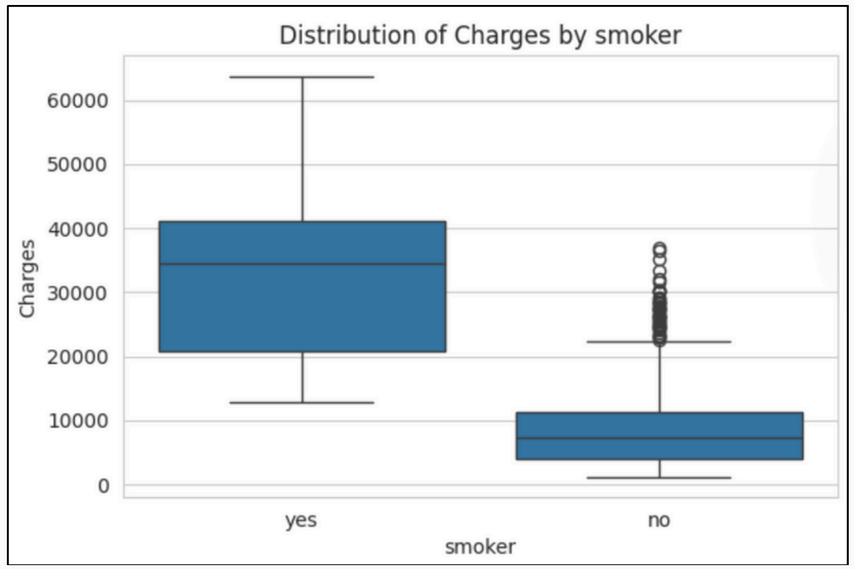
Provide interpretable insights for fairer and more transparent pricing

What the Data Tells Us

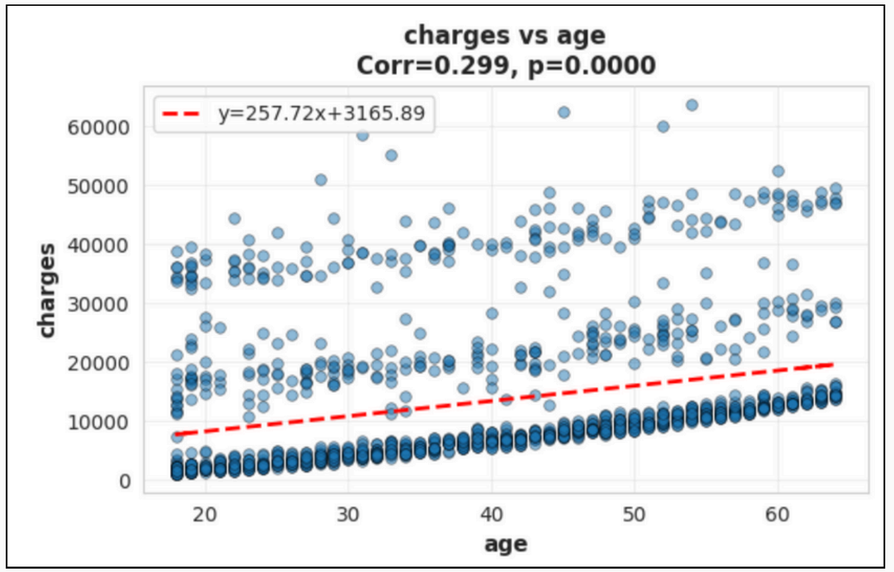


Top 3 features with strongest correlation to charges

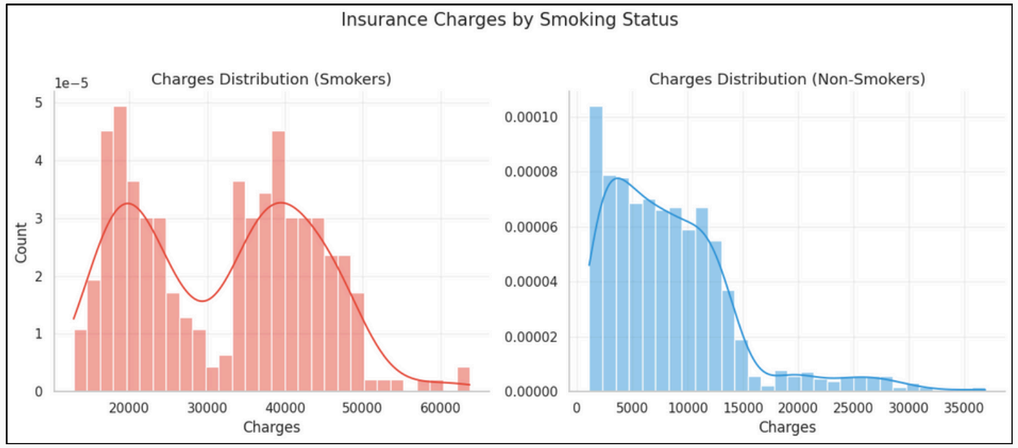
1. Smoker (0.79)
2. Age (0.30)
3. BMI (0.20)



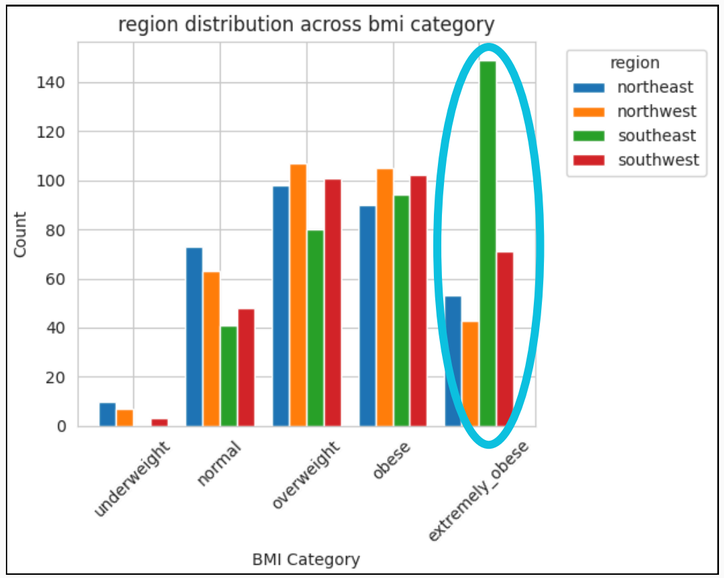
Smokers clearly have higher charges than non-smokers



Charge increases with age

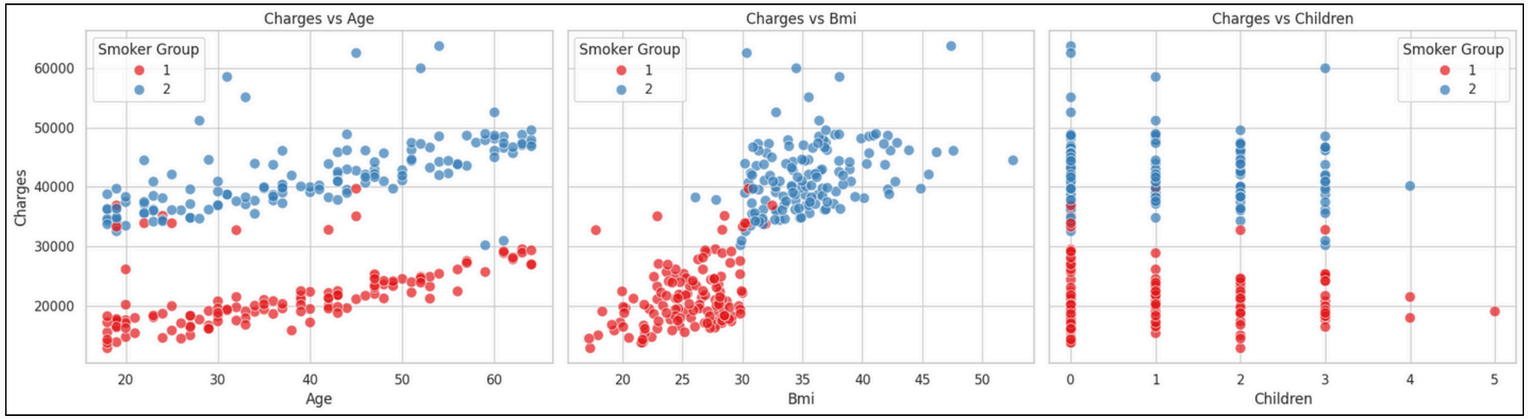


We found that there are 2 apparent subgroups among smokers.

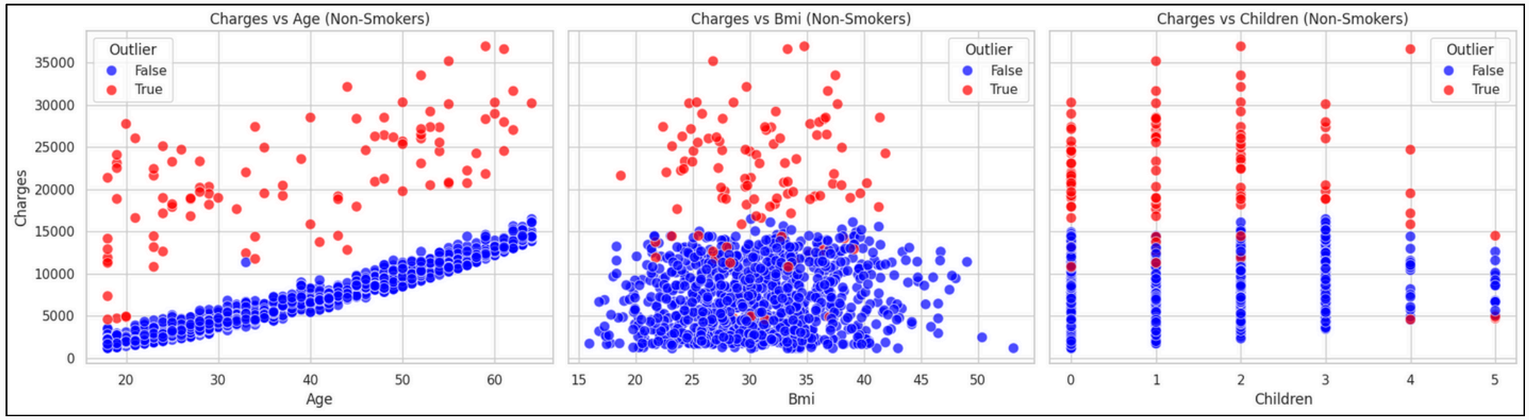


We found a higher proportion of extremely obese people in the southeast. Further investigations found that higher charges in the region are explained by smoking instead.

SMOKERS' ANALYSIS



NONSMOKERS' ANALYSIS



SUMMARY

Smokers generally incur higher medical charges than non-smokers, with two subgroups among smokers separated by BMI. Among non-smokers, there is a main group and a set of outliers, likely influenced by factors not captured in the dataset, such as medical history, lifestyle, occupation, genetics, or regional healthcare pricing differences.

Modelling Pipeline

PRE-PROCESSING

- Scaling numeric variables
- One-hot encoding categorical features
- Outlier detection and handling

MODELS TESTED

- Linear (Baseline) & Regularized models
- Tree-based ensemble models
- Custom Rule-based hybrid model

VALIDATION

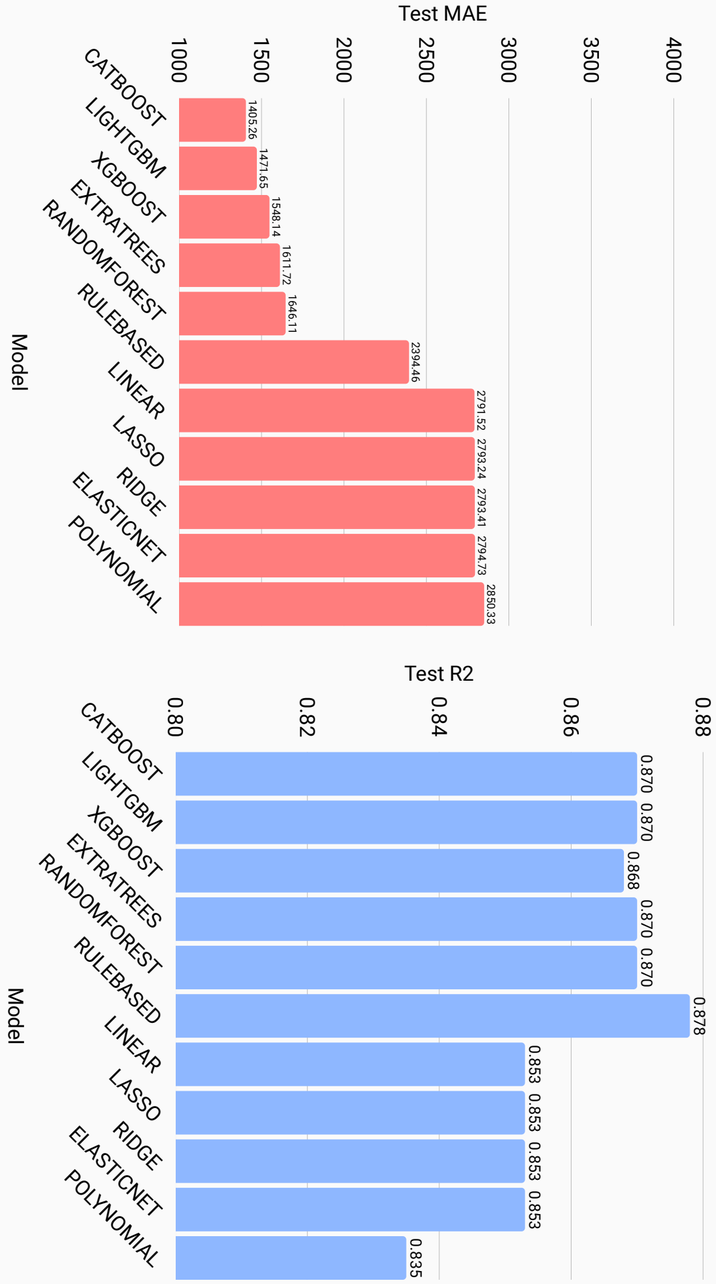
- 80/20 Train-Test Split
- 5-Fold Cross Validation

METRICS

- Mean Absolute Error (MAE)
- R^2 Score
- Variance Across Folds

Model Performances

Rank	Model	Train CV MAE	Test MAE	Test R²	MAE Gap
1	**CATBOOST**	1566.37 ± 52.28	**1405.26 ± 228.92**	0.870 ± 0.028	-161.11
2	**LIGHTGBM**	1647.14 ± 60.28	**1471.65 ± 234.88**	0.870 ± 0.027	-175.49
3	**XGBOOST**	1745.51 ± 65.78	**1548.14 ± 203.61**	0.868 ± 0.026	-197.37
4	**EXTRATREES**	1787.69 ± 45.76	**1611.72 ± 216.90**	0.870 ± 0.027	-175.97
5	**RANDOMFOREST**	1823.08 ± 60.43	**1646.11 ± 167.16**	0.870 ± 0.025	-176.97
6	**RULEBASED**	2505.36 ± 81.35	**2394.46 ± 141.93**	0.878 ± 0.021	-110.89
7	LINEAR	2957.03 ± 99.92	2791.52 ± 202.74	0.853 ± 0.034	-165.51
8	LASSO	2954.06 ± 100.03	2793.24 ± 199.83	0.853 ± 0.034	-160.82
9	RIDGE	2951.80 ± 98.44	2793.41 ± 200.03	0.853 ± 0.033	-158.38
10	ELASTICNET	2951.82 ± 98.46	2794.73 ± 200.20	0.853 ± 0.033	-157.08
11	POLYNOMIAL	3008.81 ± 84.42	2850.33 ± 179.56	0.835 ± 0.027	-158.49



Statistical Insights

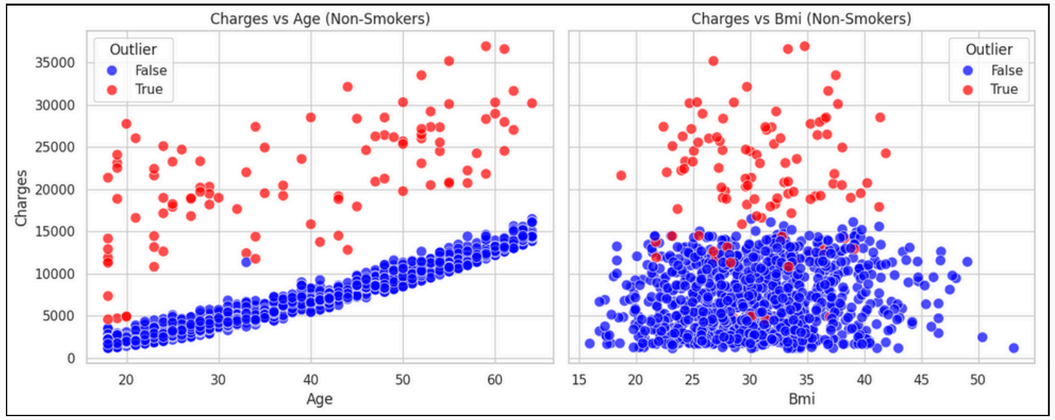
FEATURE RELEVANCE (CATBOOST)

Rank	Feature	Feature Importance (%)
1	smoker_bmi	38.55
2	smoker_age	21.94
3	smoker_yes	19.42
4	age_squared	6.09
5	bmi_squared	3.95

STATISTICAL INFERENCE

Feature importance reveals that **smoking-related** factors enable the **most accurate** predictions on medical insurance costs due to their high importance. This is supported by our EDA, where we found higher charges for smokers with higher BMI and age. The most important features relevant to **non-smokers** have significantly lower importance so predictions on cost are much **less accurate**.

MODEL LIMITATIONS



Our EDA process uncovered two clusters of points within the non-smoker sub-group. We would expect to find a feature that uses this distinction to predict charges accurately. Yet, feature importance showed no such feature. It is likely that factors uncaptured by the dataset are responsible for the outliers. This means our models will systematically perform poorly on non-smokers with an expected higher insurance cost. This problem could be overcome with a much larger dataset with more features.

Fairness Analysis

Issue	Root Cause	Impact & Mechanism	Who's Affected	Unfairness
1. Asymmetric Predictive Power	Smoking features = 80% predictive power; non-smoking features = 20%; charges for model are right-skewed	Weak features → model regresses to mean ($13K) for non-smokers → high-risk non-smokers likely undercharged	• High-risk non-smokers (benefit) • Insurance company (loses money)	**UNFAIR #1**: High-risk customers game system by paying below true cost; company absorbs $12K+ losses per customer
2. Cross-Subsidisation	Company losses from undercharging → must raise premiums to stay solvent	Premiums increase for all customers to cover high-risk non-smoker losses → healthy customers pay more	• Healthy/low-risk customers (overcharged) • Company (loses pricing integrity)	**UNFAIR #2**: Healthy customers forced to subsidize high-risk customers - violates insurance principle that premiums should reflect individual risk
3. Adverse Selection	Underpriced premiums attract high-risk non-smokers who recognize bargain rates	High-risk non-smokers disproportionately enroll → risk pool worsens → more losses → higher premiums for others → healthy customers exit	• Entire insurance pool (destabilises) • Healthy customers (priced out) • Company (death spiral)	**UNFAIR #3**: System rewards strategic gaming over health management; creates death spiral where company becomes unsustainable and low-risk customers are priced out

Challenges

Our models consistently performed better on the test set as compared to the validation sets. This implies possible selection bias.

SOLUTION

1. Test our models across different random states [42, 123, 456, 789, 999]
2. Incorporate mean and standard deviation into results evaluation

RESULTS (USING CATBOOST AS EXAMPLE)

Seed	Train CV MAE	Train CV R^2	Test MAE	Test R^2	MAE Gap	R^2 Gap
42	1599.58	0.850	1347.76	0.882	-251.82	0.032
123	1626.25	0.841	1154.44	0.911	-471.81	0.071
456	1569.23	0.855	1276.00	0.866	-293.23	0.011
789	1546.93	0.854	1499.64	0.852	-47.28	-0.002
999	1489.84	0.859	1748.45	0.839	258.61	-0.020

Mean Test MAE	1405.26
SD of Test MAE	228.92
Final Test MAE	1405.26 ± 228.92

Running more statistical checks, we deduced that the largest factor is that more data points are used to train the model for predicting on the test set than that for predicting on the validation set. The generalisation improved, leading to lower average test MAE. This improvement was significant and consistent as the dataset is small.

From Insights to Impact: Practical Relevance and Next Steps

RELEVANCE

1) Insights reflect real-world insurance trends — smoking, BMI, and age are key drivers of medical insurance charges.

2) These findings support data-driven cost estimation, helping insurers and policymakers improve fairness, transparency, and resource allocation in healthcare insurance planning

RECOMMENDATIONS

1) **Enhance feature diversity**:

   Broaden data scope to include **medical history, exercise habits, diet, stress factors** and more to better explain outliers and differences among non-smokers' medical insurance charges

SCALABILITY

1) CatBoost framework is adaptable to larger and more complex datasets, supporting periodic retraining as new data streams in

   Model pipeline can integrate into insurer systems for automated risk scoring and medical insurance estimation.

Thank you

Appendix

HTTPS://COLAB.RESEARCH.GOOGLE.COM/DRIVE/1YDY_DD85WWFSCKOU-3YQX6EW_N_9S0CS?USP=SHARING