# Insurance Loss Analytics

DSO 530 Group Project

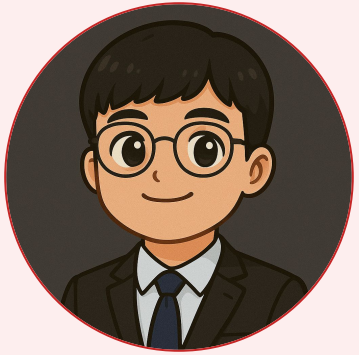
# Team Members

Phuong Vu

Shreya

Amna Gul

Anh Tuyet (Snowy) To



Chin-Wei (Wilson) Chang

Seungyun (Freddie) Lee

# Project Overview

## Industry Context

One of the main challenges in the insurance industry is setting the **right premium** for policyholders. **Mispricing** can cause **adverse selection** leading to **financial loss** .

## Problem Statement

Insurers need to **accurately predict** expected **losses** and **adjust premiums** accordingly. Accurate losses prediction will help improve premium **pricing strategies** , reduce the risk of adverse selection, and enhance **overall portfolio profitability** .

# Project Goal

Build **predictive models** to help insurance companies **set fair and profitable premiums**

# Importance & Motivation

Reduce underwriting risk and support portfolio health

Promote fairness and efficiency in pricing using **Tweedie-based ML models**

# Stakeholders

**Policyholders**
Seek transparent and justifiable premiums

**Policymakers**
Ensure market fairness and prevent discrimination

**Insurance Companies**
Need to maintain profitability

# Task 1

## Regression

**Predict:** how big the claim is

**LC (Loss Cost per Exposure Unit)** = X.15 / X.16

- Measure average cost per claim for a policyholder

**HALC (Historically Adjusted Loss Cost)** = LC × X.18

- Adjust LC to reflect how frequently claims occur historically

**Business Perspective:**

Determine **financial impact** when that risk materializes

# Task 2

## Classification

**Predict:** the probability of any claim

**Claim Status (CS)** :

- 1 = claim made
- 0 = no claim

**Business Perspective:**

Assess **risk of any claim**

## Dataset Overview

**Training Data**

- 37,451 policy records
- 28 variables (X.1–X.28)

**Feature Types**

- **Numerical** : X.8–X.12, X.14–X.18, X.22–X.26, X.28
- **Categorical:** X.7, X.13, X.19–X.21, X.27
- **Datetime:** X.2–X.6

**Test Data**

- 15,787 policy records
- Without target values (X.15–X.18)

### Datetime Conversion

- Convert **policy start date** , **renewal dates** , **date of birth** , and **license issuance date** to datetime.

### Feature Engineering

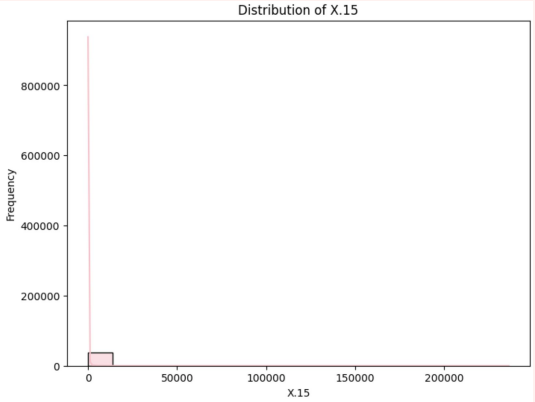
- **age** = Age of the policyholder in years
- **policy_duration** = Days between policy start and last renewal
- **driver_experience** = Years since driver's license was issued
- **Vehicle_age** = Age of vehicle as of 2019

### Columns Dropped

- **X.1**: A unique identifier for each policy, which doesn't provide much useful information and could introduce unwanted noise.
- **X.2–X.6 and X.22:** We already extracted useful components using feature engineering.

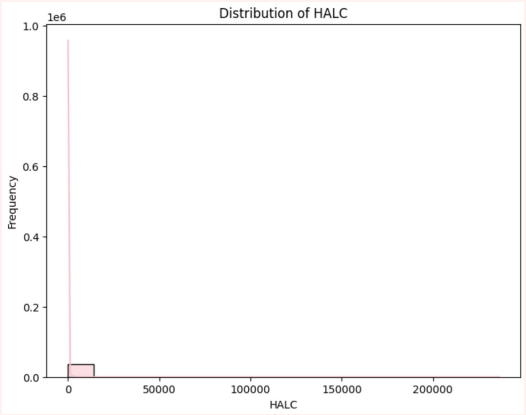
# Distributions of Numerical Features (Used for LC, HALC, CS Prediction)

## X.15–Total Claim Cost



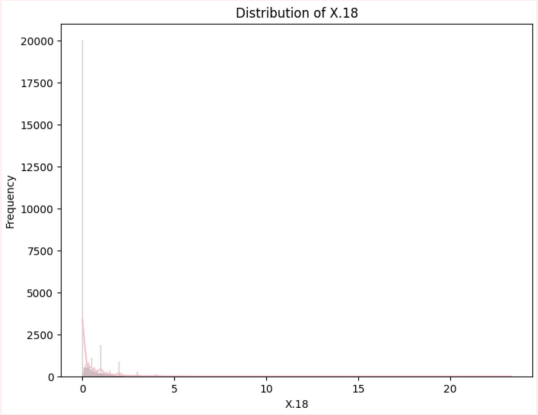

- Highly **right-skewed** with many **zeros** – most policyholders didn't file a claim.
- Few records show **extreme claim costs** , highlighting **rare but large incidents** .

## X.15 / X.16–Loss Cost




- Extreme **spike at zero** and long tail – many low-cost claims, few expensive ones.
- Outliers with **very high per-claim costs** , possibly from rare high-payout policies.

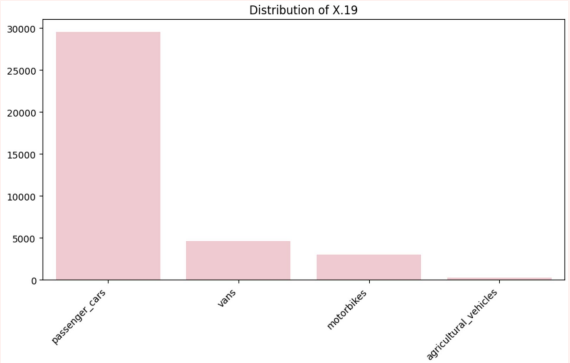
## X.18–Claim Frequency




- **Heavily right-skewed** , with most values close to zero.
- Claims occur **infrequently** across most policies.

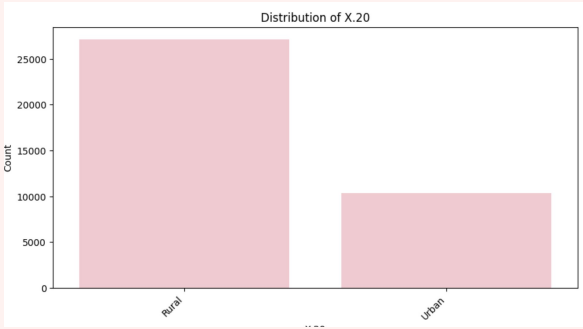
# Distributions of Categorical Features

## X.19–Vehicle Type



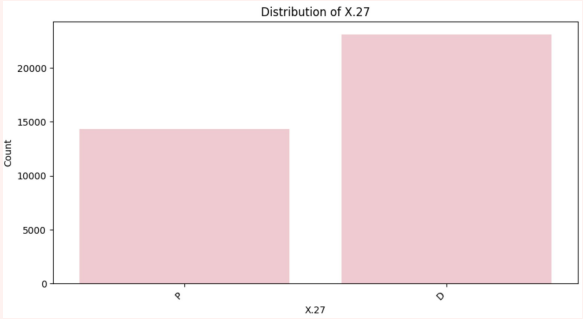

- **Passenger cars** make up the vast majority of insured vehicles.
- **Vans** and **motorbikes** represent much smaller segments.
- Very few **agricultural vehicles** , indicating they may have limited impact on the model.

## X.20–Region Type




- The dataset is **dominated by rural customers** , with urban policyholders being a smaller group.

## X.27–Fuel Type




- Majority of vehicles run on **diesel** , but **petrol** still forms a significant portion.
- Class balance is relatively better than other categorical features.

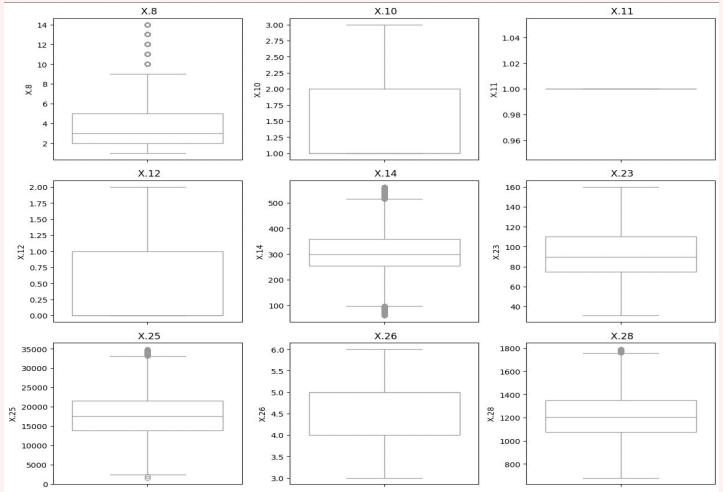
# Handling Outliers

## Regression

Thresholds Applied

Outliers beyond 1.5 × IQR from the 1st and 3rd quartiles

~36% of rows were removed in total




## Classification

Thresholds Applied

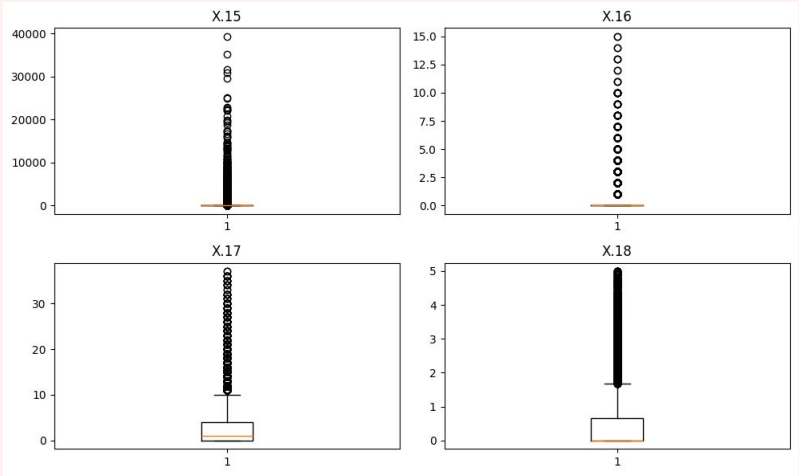
X.15 > 40000

X.16 > 15

X.17 > 40

X.18 > 10

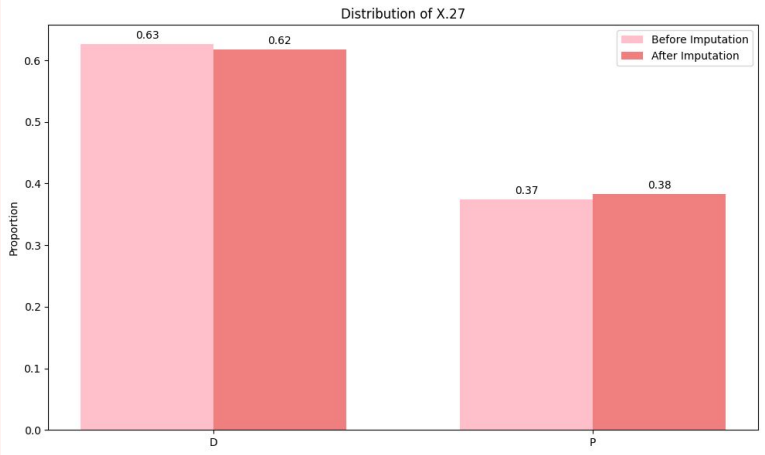
Less than 0.1% of rows were removed in total



# Handling Missing Values in X.27 (Fuel Type)

## Approach

- Imputed using supervised classification models
  - Features used: X.19, X.23, X.24, X.25, X.28
- Final imputed values were integrated back into the training dataset




## Comparison

| Model | Decision Tree / Random Forest | XGBoost | LightGBM |
|---|---|---|---|
| Accuracy | 99.76% | 96.31% | 94.7% |
| Recall (D / P) | 1.00 / 1.00 | 0.98 / 0.94 | 0.97 / 0.91 |
| Observation | Overfit to training set, poor recall on minority class P | Best balance between performance and generalization; strong recall on both classes, especially minority P | Slight underfitting, missed both classes more often |

# Features Used & Correlation Heatmap

## Columns with correlation > |0.8|

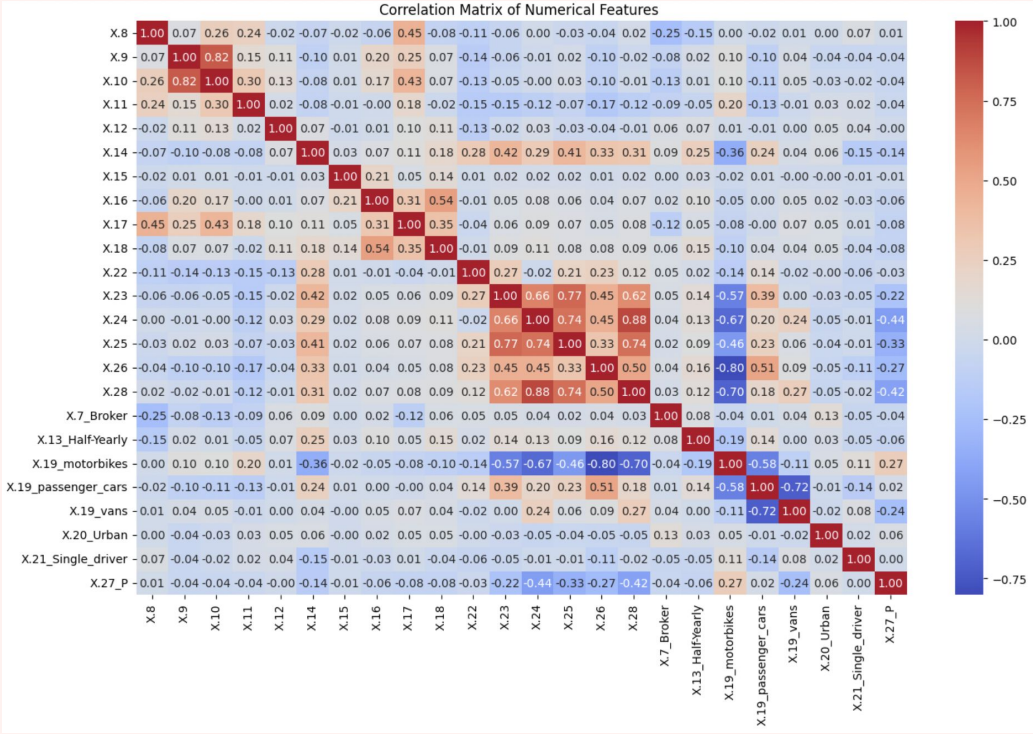
- 'X.9' and 'X.10' = 0.82
- 'X.24' and 'X.28' = 0.88
- 'X.26' and 'X.19_motorbikes' = -0.8

## Columns dropped

Drop **'X.9', 'X.24', 'X.19_motorbikes'** columns apart from date and year columns

## Final features used

- **Numerical:** 'X.8', 'X.10', 'X.11', 'X.12', 'X.14', 'X.23', 'X.25', 'X.26', 'X.28'
- **Categorical:** 'X.7_Broker', 'X.13_Half-Yearly', 'X.19_passenger_cars', 'X.19_vans', 'X.20_Urban', 'X.21_Single_driver', 'X.27_P'
- **Prediction:** 'LC', 'HALC', 'CS'



## Regression *

Prediction Target: LC & HALC
Evaluation Metric: RMSE

| Model Type | Model | LC | HALC |
|---|---|---|---|
| Linear | Tweedie Regression | 573.67 | 1078.13 |
| Non-Linear | LightGBM | 571.79 | 1072.38 |
| | XGBoost | 574.04 | 932.68 |
| | **Neural Network** | **458.92** | **718.54** |

*All models trained using Tweedie Loss Function

## Classification

Prediction Target: CS
Evaluation Metric: ROC-AUC

| Model Type | Model | ROC-AUC Score |
|---|---|---|
| Linear | Logistic Regression | 0.74 |
| | **XGBoost** | **0.85** |
| Non-Linear | Random Forest | 0.82 |
| | Neural Network | 0.71 |

# Business Understandings

## Business Perspective

- **Prediction accuracy** is more critical for setting proper premiums.
- Accurate risk prediction ensures **profitability** and protects **against adverse selection** .
- The cost of inaccurate predictions outweighs the need for full model transparency.

## Key Challenges

- **Highly skewed and zero-inflated data:** Rare, costly claims challenge prediction.
- **Complexity vs. Interpretability:** Models must balance accuracy and transparency.
- **Generalization to new customers:** Models must adapt to unseen policyholders.

## Variable Selection

- Driver characteristics
- Vehicle attributes
- Policy details
- Location information

## Applicability

Car insurance model cannot be directly used for life insurance.

- Different risk objects (vehicle vs human life)
- Different predictive features

# THANK YOU!

# Business Implications

## Business Perspective

- **Prediction accuracy** is more critical for setting proper premiums.
- Accurate risk prediction ensures **profitability** and protects **against adverse selection** .
- The cost of inaccurate predictions outweighs the need for full model transparency.

## Key Challenges

- **Highly skewed and zero-inflated data:** Rare, costly claims challenge prediction.
- **Complexity vs. Interpretability:** Models must balance accuracy and transparency.
- **Generalization to new customers:** Models must adapt to unseen policyholders.

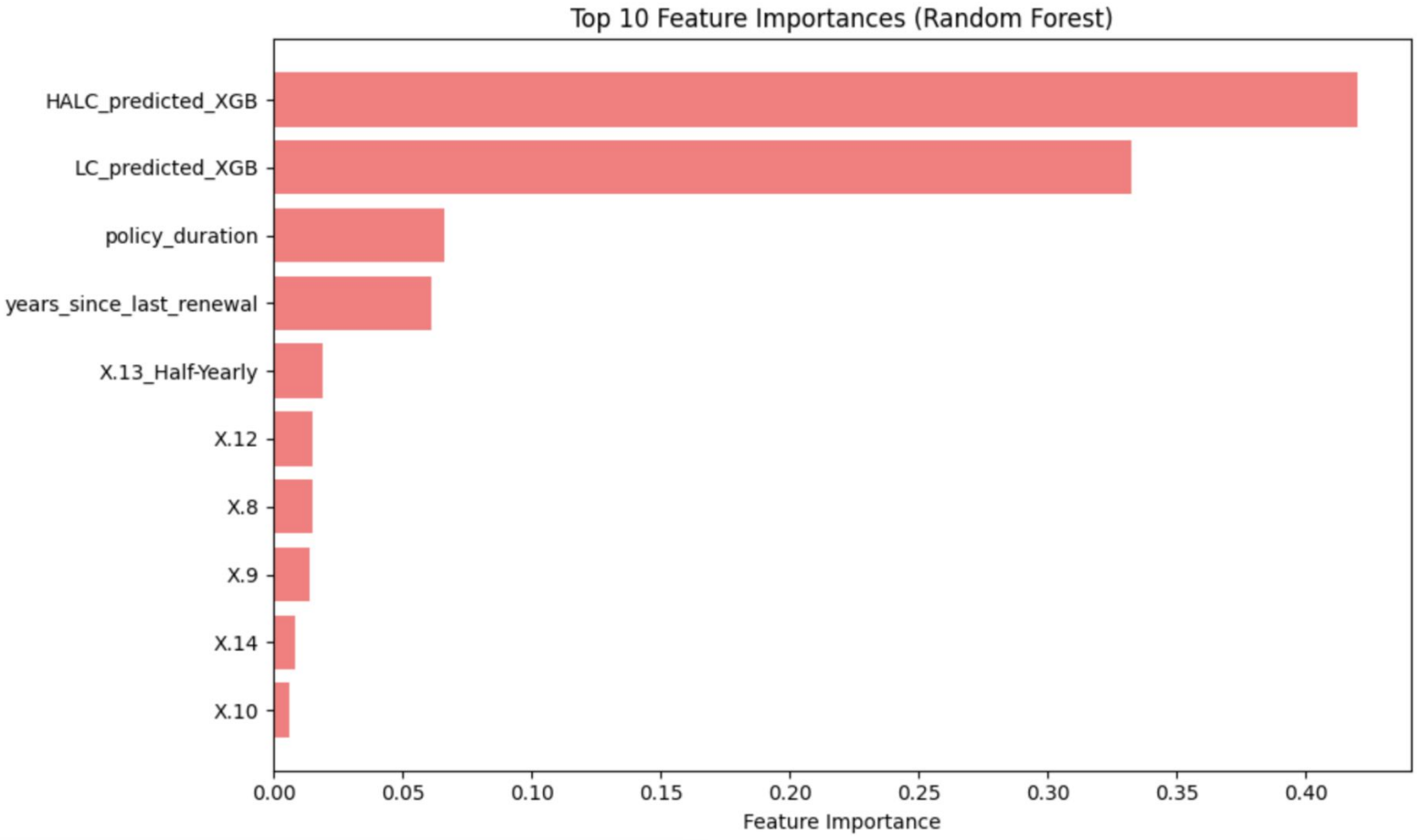
## Variable Selection

- Driver characteristics
- Vehicle attributes
- Policy details
- Location information

## Applicability

Car insurance model cannot be directly used for life insurance.

- Different risk objects (vehicle vs human life)
- Different predictive features


Top 10 Feature Importances (Random Forest)
HALC_predicted_XGB
LC_predicted_XGB
policy_duration
years_since_last_renewal
X.13_Half-Yearly
X.12
X.8
X.9
X.14
X.10
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Feature Importance

# Business Understanding:

USC Marshall School of Business

## Business Perspective

Prediction accuracy is more critical than interpretability. Setting premiums based on expected risk demands highly accurate predictions to ensure profitability and avoid adverse selection. For the insurance business, the cost of inaccurate predictions outweighs the need for full model transparency

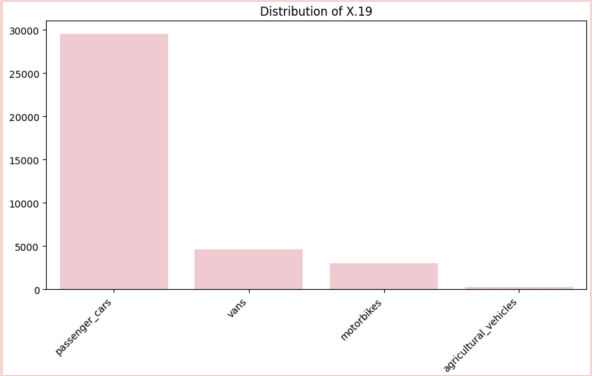
## Key Challenges

## Applicability

# Exploratory Data Analysis (EDA)

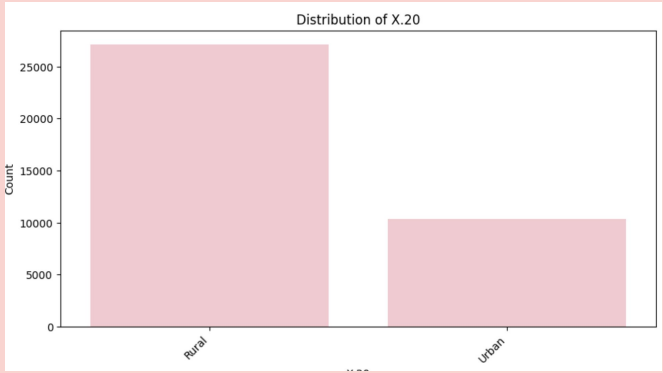
## Categorical- Boxplots

### X.19 — Vehicle Type




- **Passenger cars** make up the vast majority of insured vehicles.
-**Vans and motorbikes** represent much smaller segments.
-Very few **agricultural vehicles** , indicating they may have limited impact on the model.

### X.20 — Region Type




-The dataset is **dominated by rural customers** , with urban policyholders being a smaller group

# Introduction

USC Marshall School of Business

## Goal

Build **predictive models** to help insurance companies **set fair and profitable premiums**

## Problem Statement

Insurance companies face challenges in **setting accurate premiums**. Overcharging low-risk drivers or underpricing high-risk ones can lead to **adverse selection** and financial loss.

## Industry Context

In the **automobile insurance industry**, claim distributions are **highly skewed** with many zero claims. Traditional models struggle to handle this, especially when aiming for fair and profitable pricing.

## Stakeholders

**Policyholders**: Seek transparent and justifiable premiums
**Policymakers**: Ensure market fairness and prevent discrimination
**Insurance Companies**: Need to maintain **profitability**

## Importance & Motivation

Reduces underwriting risk and supports portfolio health
Promotes fairness and efficiency in pricing using **Tweedie-based ML models**

# Exploratory Data Analysis (EDA) | Dataset Overview

USC Marshall School of Business

## Dataset Overview

**Training Data**

- 37,451 records
- Variables: X.1 to X.28

**Feature Types**

- **Numerical**: X.8–X.12, X.14–X.18, X.22–X.26, X.28
- **Categorical:** X.7, X.13, X.19–X.21, X.27
- **Datetime:** X.2–X.6

## Datetime Conversion

- Convert policy start date, renewal dates, date of birth, and license issuance date to date time.

## Feature Engineering

- age = Age of the policyholder in years
- policy_duration = Days between policy start and last renewal
- driver_experience = Years since driver's license was issued
- Vehicle_age = Age of vehicle as of 2019

## Columns dropped

- **X.1**: a unique identifier for each policy. It doesn't provide any useful information for our modeling tasks and could introduce unwanted noise.
- **X.2-X.6 and X.22:** We already extracted useful components using feature engineering.

# Final Results |

USC Marshall School of Business

## Regression

- Neutral Network

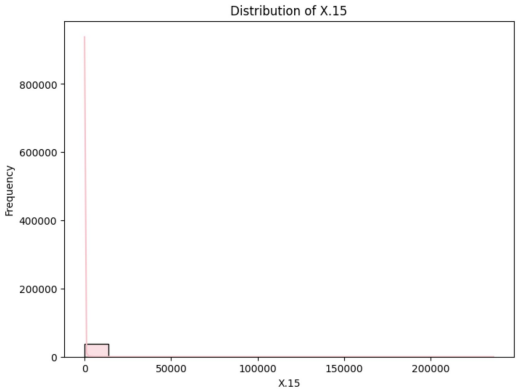
## Classification

- Entropy for best model

# Exploratory Data Analysis (EDA) | Numerical- Histogram and Distribution

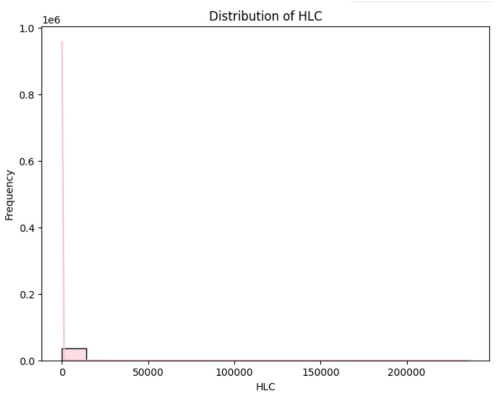
USC Marshall School of Business

## X.15 — Total Claim Cost



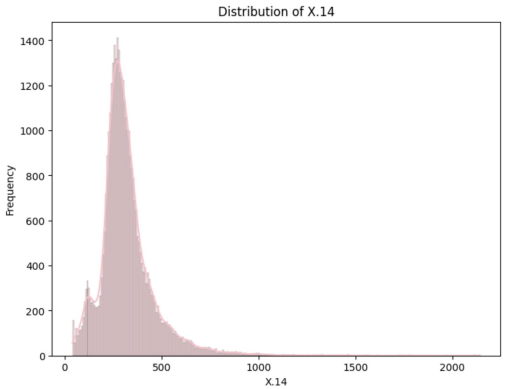

- Highly **right-skewed** with a large concentration at **zero** → most policyholders didn't file a claim.
- A small number of records have **extremely high claim costs**, showing the **rarity but impact** of large incidents.

## X.15 / X.16 — Loss Cost




- Extreme **spikes at zero** and a long tail → many low-cost claims, few very expensive ones.
- Some outliers with **very high per-claim costs**, possibly from policies with few claims but very high payouts.

## X.14 — Net Premium



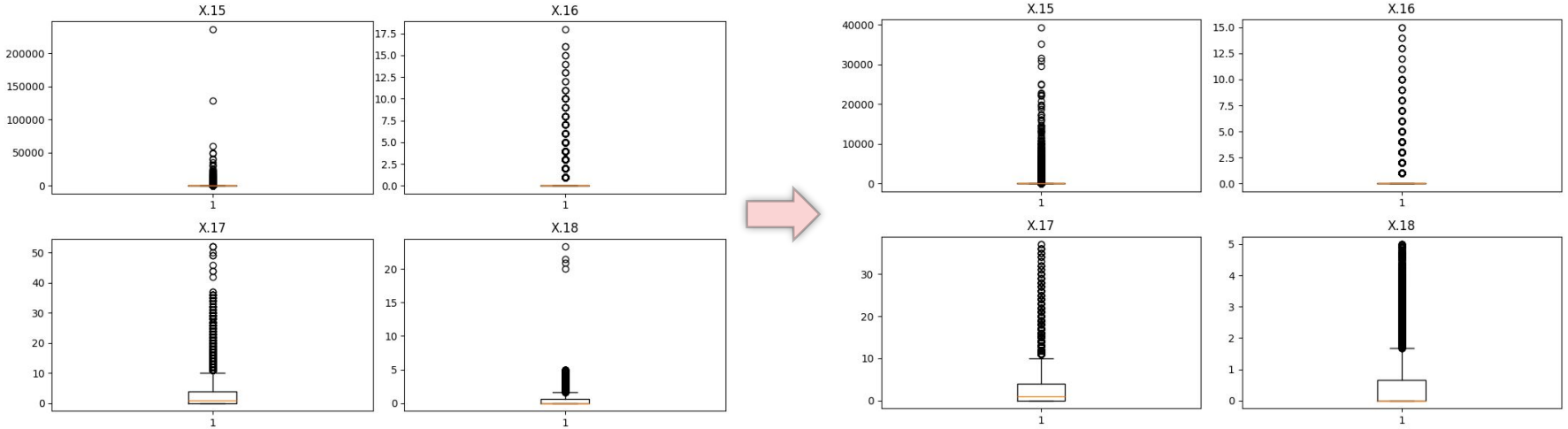

- Distribution is **positively skewed** but more stable than claim-related variables.
- Most policyholders pay around **300–400 units** in premiums, with a clear peak.

# Exploratory Data Analysis (EDA) | Outliers (X.15-X.18)

USC Marshall School of Business

## Outliers

- Thresholds Applied:
  - X.15 > 100,000
  - X.16 > 15
  - X.17 > 40
  - X.18 > 10
- Less than 0.1% of rows were removed in total

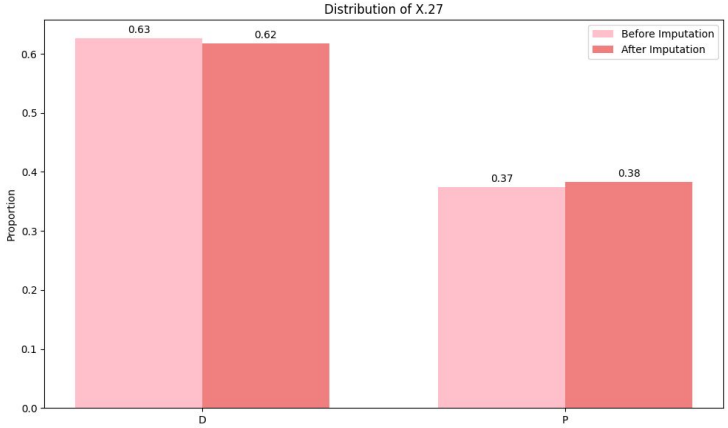


# Exploratory Data Analysis (EDA) | Fuel Type (X.27)

USC Marshall School of Business

## Handling Missing Values

- Only **X.27 (fuel type)** had missing values
- Imputed using supervised classification models
  - Features used: X.19, X.23, X.24, X.25, X.28
- **XGBoost** was selected for its balanced performance and generalization
- Final imputed values were integrated back into the training dataset

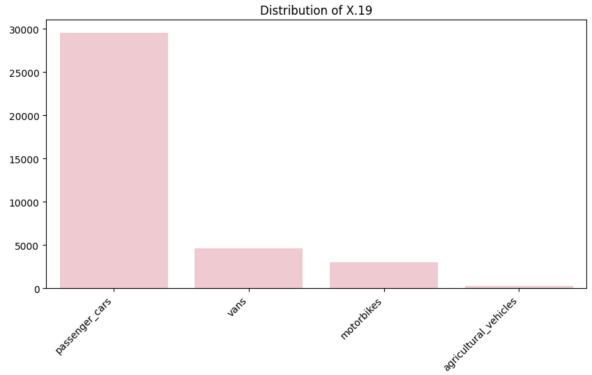



| Model | Accuracy | Recall (D / P) | Observation |
|---|---|---|---|
| Decision Tree / Random Forest | 99.76% | 1.00 / 1.00 | Overfit to training set, poor recall on minority class P |
| **XGBoost** ✅ | 96.31% | 0.98 / 0.94 | Best balance between performance and generalization; strong recall on both classes, especially minority P |
| LightGBM | 94.7% | 0.97 / 0.91 | Slight underfitting, missed both classes more often |

# Exploratory Data Analysis (EDA) | Categorical- Boxplots

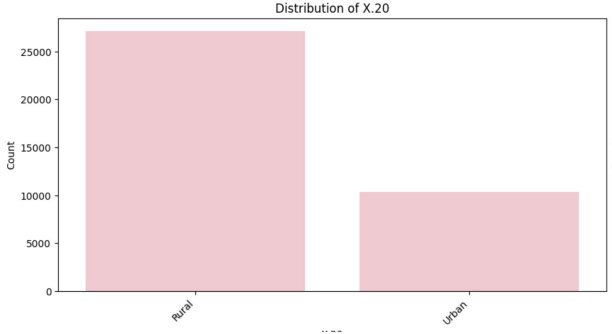
USC Marshall School of Business

## X.19 — Vehicle Type




- **Passenger cars** make up the vast majority of insured vehicles.
- **Vans** and **motorbikes** represent much smaller segments.
- Very few **agricultural vehicles**, indicating they may have limited impact on the model.

## X.20 — Region Type




- The dataset is **dominated by rural customers**, with urban policyholders being a smaller group

## X.27 — Fuel Type


- Majority of vehicles run on **diesel**, but **petrol** still forms a significant portion.
- Class balance is relatively better than other categorical features.

| Regression - RMSE | | | |
|---|---|---|---|
| **Model Type** | **Models** | **LC** | **HALC** |
| Linear | Tweedie Regression | 573.67 | 1078.13 |
| Non-Linear | LightGBM | 571.79 | 1072.38 |
| | XGBoost | 574.04 | 932.68 |
| | **Neural Network** | **458.92** | **718.54** |

| Classification- CS | | |
|---|---|---|
| **Model Type** | **Models** | **ROC-AUC Score** |
| Linear | Logistic Regression | 0.74 |
| Non-Linear | **XGBoost** | **0.85** |
| | Random Forest | 0.82 |
| | SVM | 0.50 |

# Premium Active Wear Industry

## Industry Intersection

Athletic Apparel, Fashion, Wellness

## Athleisure Clothing

High-Quality, Performance-Oriented, Daily

## Product Emphasis

Technical Materials, Elevated Aesthetics, Lifestyle Branding

## Industry Lifecycle: Growth Stage

U.S. Market Valuation Projection: $67B in 2022 $85B in 2025

Frequent new entrants launching differentiated products

Rapid expansion of direct-to-customer channels

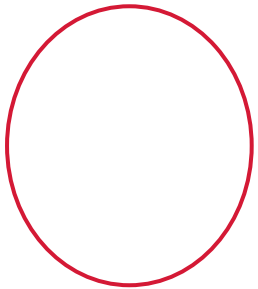
## Customer Demographics

Age Range: 18-35

Middle to high income

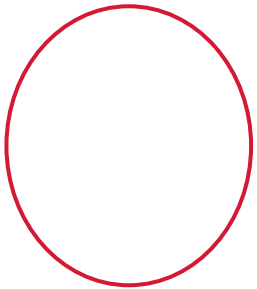
Prioritize health and wellness

## Customer Demographics



**Ocean (Age: 32, Female)**

- Successful, health-conscious
- Enjoys travelling
- 1.5-hour workout per day



**Duke (Age: 35, Male)**

- “Athletic opportunist”
- Surfing, snowboarding
- Willing to pay for quality

# Table of contents

# 01. Objectives

You can enter a subtitle here if you need it

# Introduction

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than the Moon. The planet's name has nothing to do with the liquid metal, since it was named after the Roman messenger god, Mercury

“This is a quote, words full of wisdom that someone important said and can make the reader get inspired.”

**—Someone Famous**

# Study objectives

## Mars

Mars is full of iron oxide dust, which gives the planet its reddish cast

## Venus

Venus has a beautiful name and is the second planet from the Sun

## Mercury

Mercury is the closest planet to the Sun and the smallest one of them all

# Schedule

| | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| Week 1 | | | | 1 | 2 | 3 | 4 |
| Week 2 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| Week 3 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| Week 4 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| Week 5 | 26 | 27 | 28 | 29 | 30 | | |

# Our objectives

## Mars

Despite being red, Mars is actually a cold place

## Mercury

Mercury is the closest planet to the Sun

## Venus

Venus has a beautiful name, but it’s hot

## Jupiter

Jupiter is a gas giant and the biggest planet

## Neptune

Neptune is the farthest planet from the Sun

## Saturn

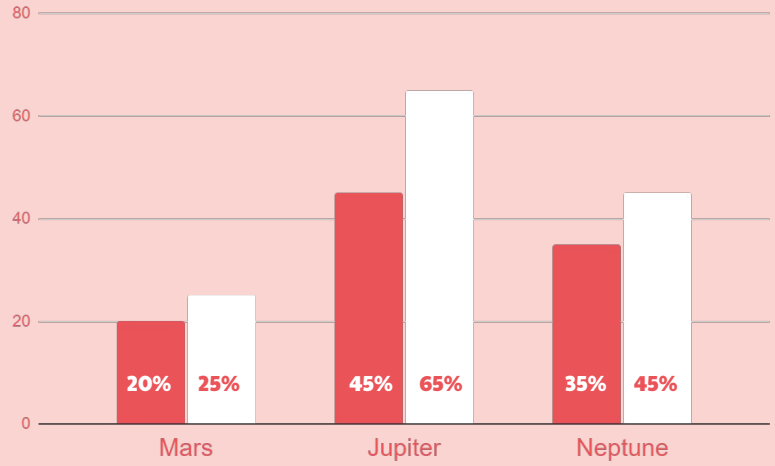
It’s composed of hydrogen and also helium

# A picture is worth a thousand words

# Table

| | Test 1 | Test 2 | Test 3 | Test 4 |
|---|---|---|---|---|
| Describe here your metrics 1 | ✗ | ✓ | ✗ | ✓ |
| Describe here your metrics 2 | ✗ | ✓ | ✗ | ✓ |
| Describe here your metrics 3 | ✗ | ✓ | ✗ | ✓ |

# Results analysis




**60%**

Mercury is the smallest and innermost planet in the entire Solar System

To modify this graph, click on it, follow the link, change the data and paste the resulting graph here, replacing this one

# Conclusions

**01.**

The planet’s name has nothing to do with the liquid metal since it was named after the Roman messenger god, Mercury

Venus has a beautiful name and is the second planet from the Sun. It’s terribly hot—even hotter than Mercury

**02.**

# About me

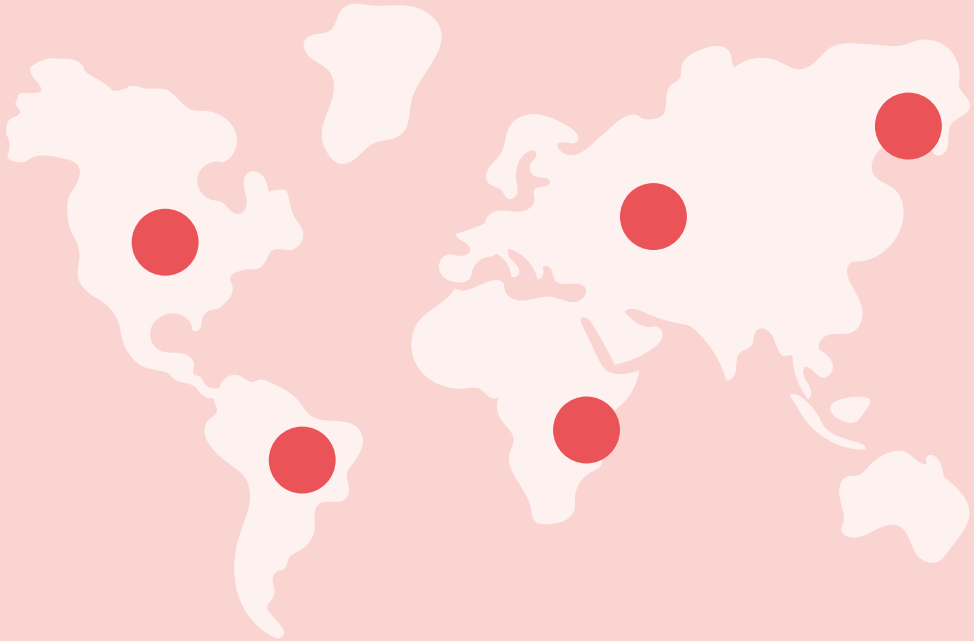
## Laura Smith

Mercury is the smallest and innermost planet in the entire Solar System

# This is a map

## Mars

The planet’s name has nothing to do with the liquid metal, since it was named after the Roman messenger god, Mercury



# Bibliographical references


- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher
- AUTHOR (YEAR). *Title of the publication*. Publisher

# Results presentation

It’s the sixth planet from the Sun and the second-largest

# 80

Mercury is the closest planet to the Sun

# 4852

Venus has a beautiful name

# 250M

It’s the sixth planet from the Sun

# Our timing

01.
## January
The closest planet to the Sun

02.
## February
Venus has a beautiful name

03.
## March
A gas giant and the biggest planet

04.
## April
Pluto is now a dwarf planet

05.
## May
Earth is the planet where we live on

06.
## Jun
Neptune is very far away from Earth

# 250,000

Jupiter is a gas giant and the biggest planet

# Thanks

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourcompany.com

# Alternative resources

## Photos

- Group of women spending time together on their phones
- Group of women having fun together
- Group of women holding each other
- Group of women looking together through a laptop
- Group of friends looking on a phone with copy space
- Group of best friends posing

# Resources

## Photos

- Group of women posing in a fun way
- Portrait of a confident young businesswoman sitting on wheelchair using laptop in the office
- Man showing something on a laptop to his coworker

# Instructions for use (free users)

In order to use this template, you must credit **Slidesgo** by keeping the Thanks slide.

**You are allowed to:**

- Modify this template.
- Use it for both personal and commercial purposes.

**You are not allowed to:**

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

# Instructions for use (premium users)

In order to use this template, you must be a Premium user on **Slidesgo**.

**You are allowed to:**

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

**You are not allowed to:**

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

https://slidesgo.com/faqs and https://slidesgo.com/slidesgo-school

# Fonts & colors used

This presentation has been made using the following fonts:

**Paytone One**
(https://fonts.google.com/specimen/Paytone+One)

**Questrial**
(https://fonts.google.com/specimen/Questrial)

| | | | | |
|---|---|---|---|---|
| #191919 | #ffffff | #595959 | #eeeeee | #f9d4d0 |
| #ea5458 | #fef2f0 | #ffffff | #f9d4d0 | #ea5458 |

# Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out How it Works.

Pana

Amico

Bro

Rafiki

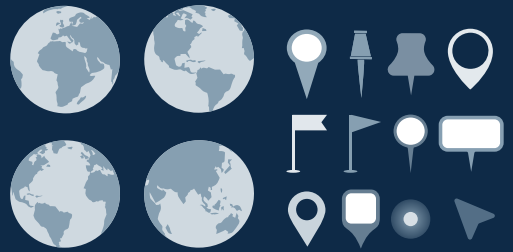
Cuate

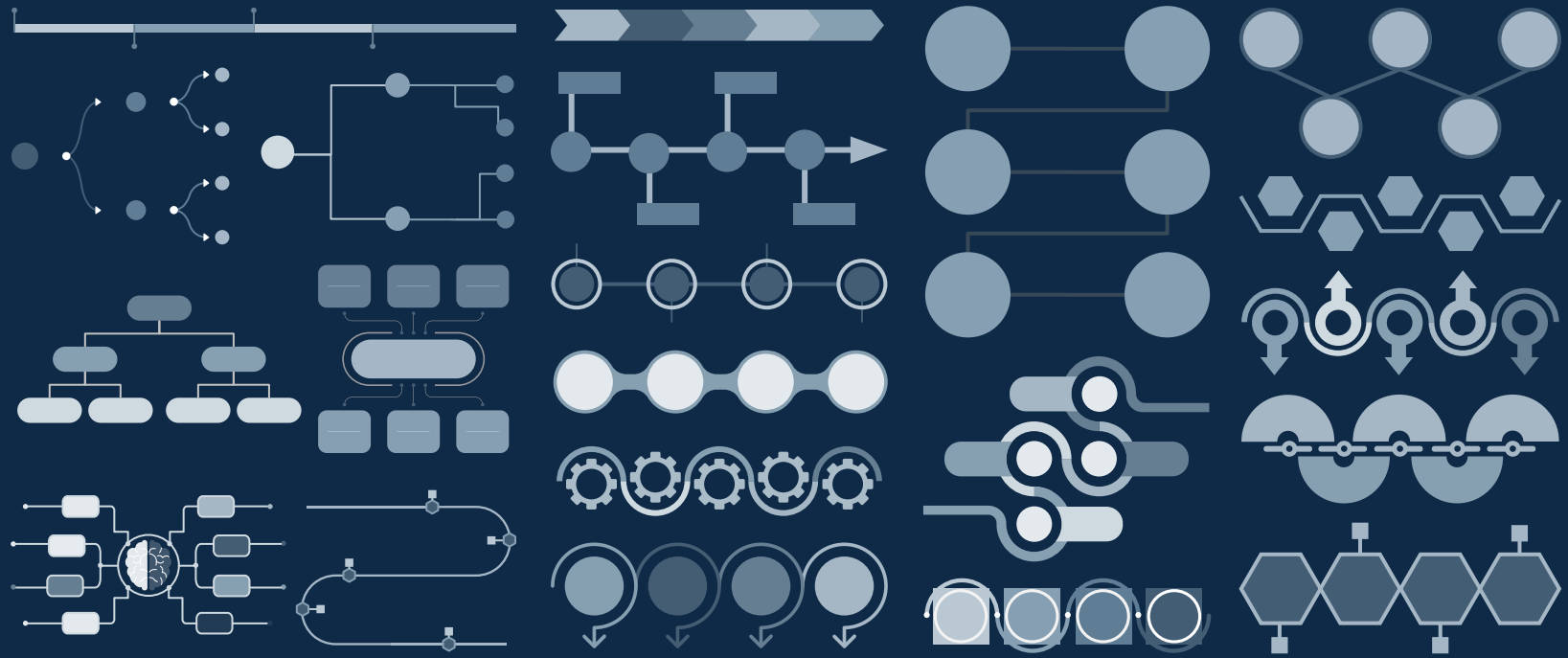
# Use our editable graphic resources...

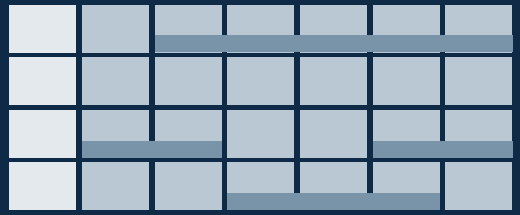
You can easily resize these resources without losing quality. To change the color, just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want. Group the resource again when you're done. You can also look for more infographics on Slidesgo.

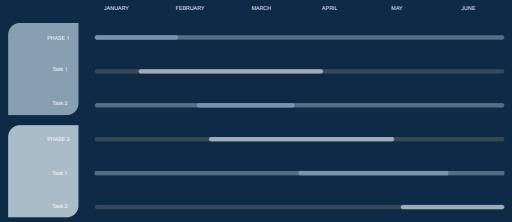


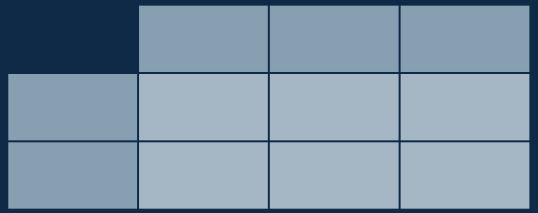


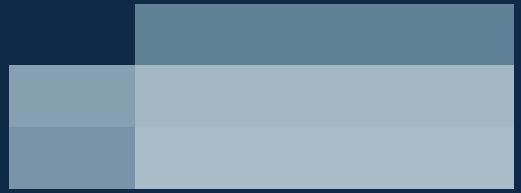



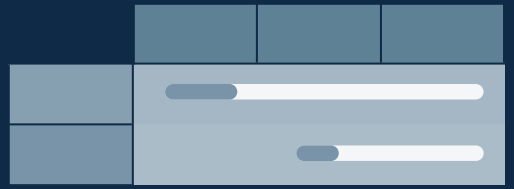


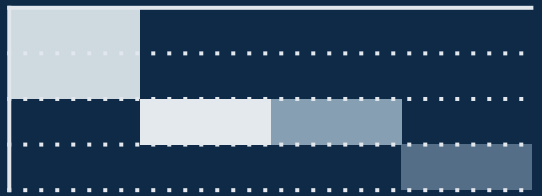


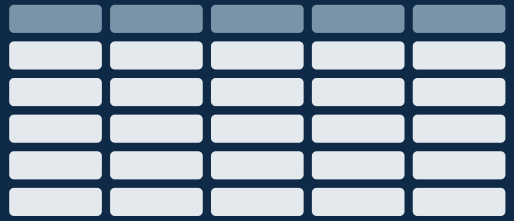

JANUARY
FEBRUARY
MARCH
APRIL
MAY
JUNE
PHASE 1
Task 1
Task 2
PHASE 2
Task 1
Task 2

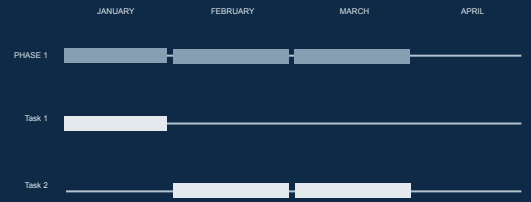


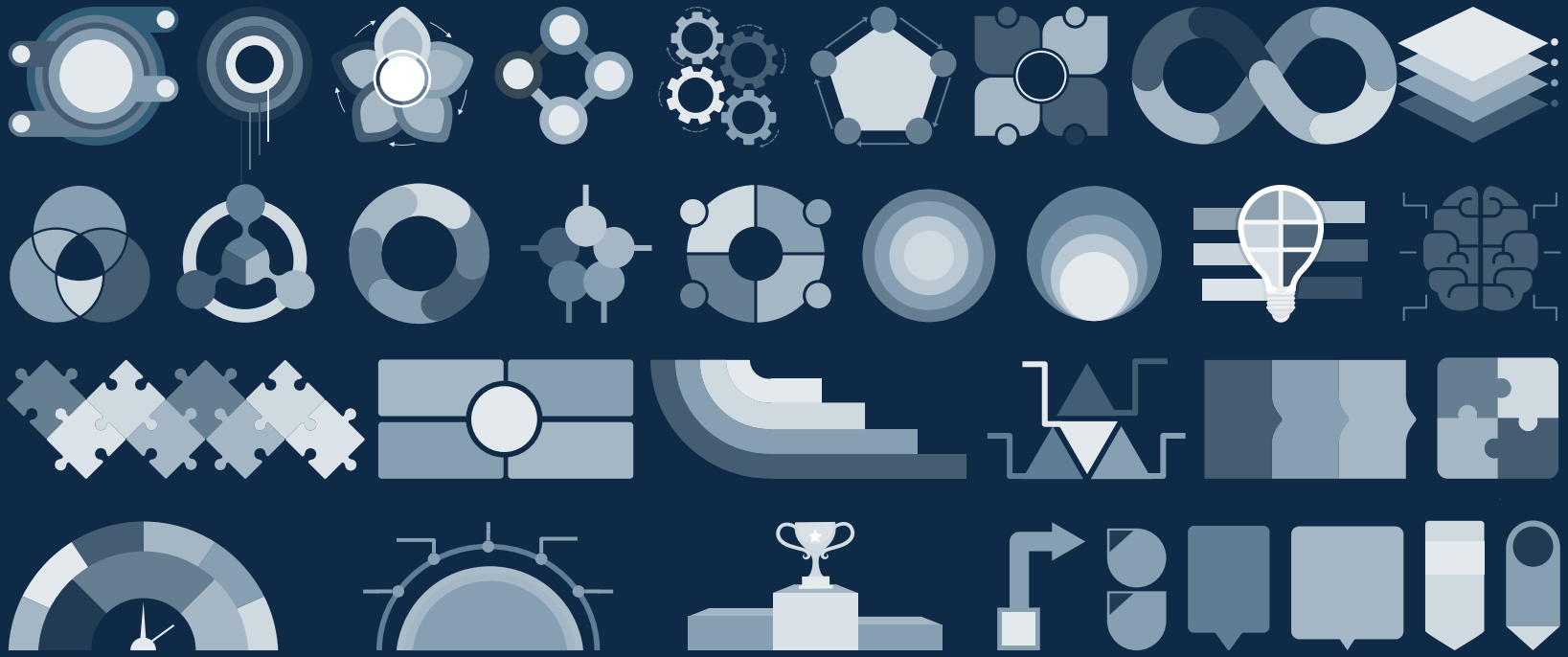


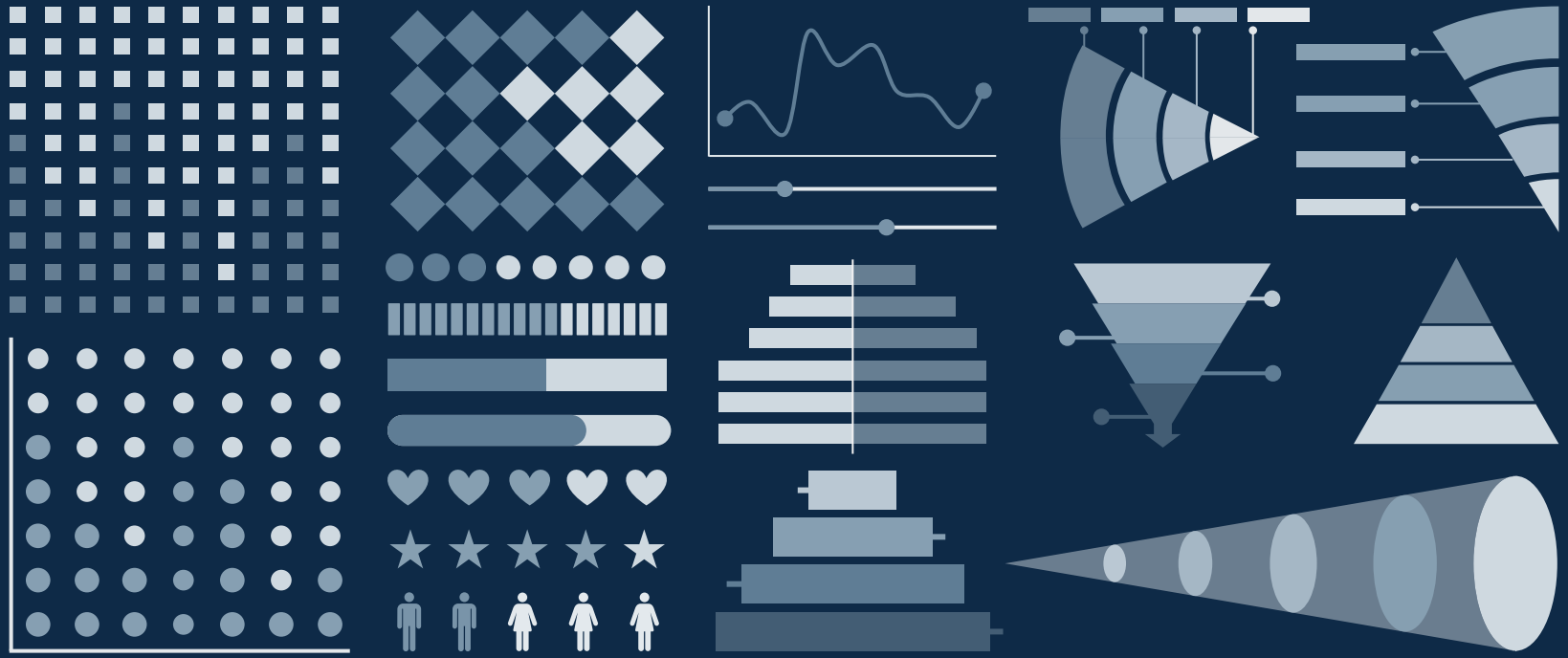







JANUARY
FEBRUARY
MARCH
APRIL
PHASE 1
Task 1
Task 2





# ...and our sets of editable icons

You can resize these icons without losing quality.
You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.
In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.

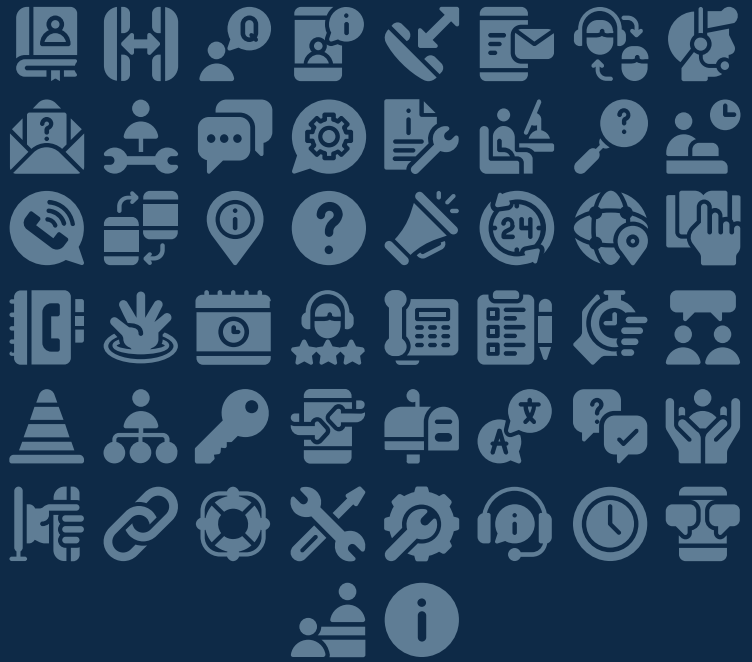
## Educational Icons

## Medical Icons

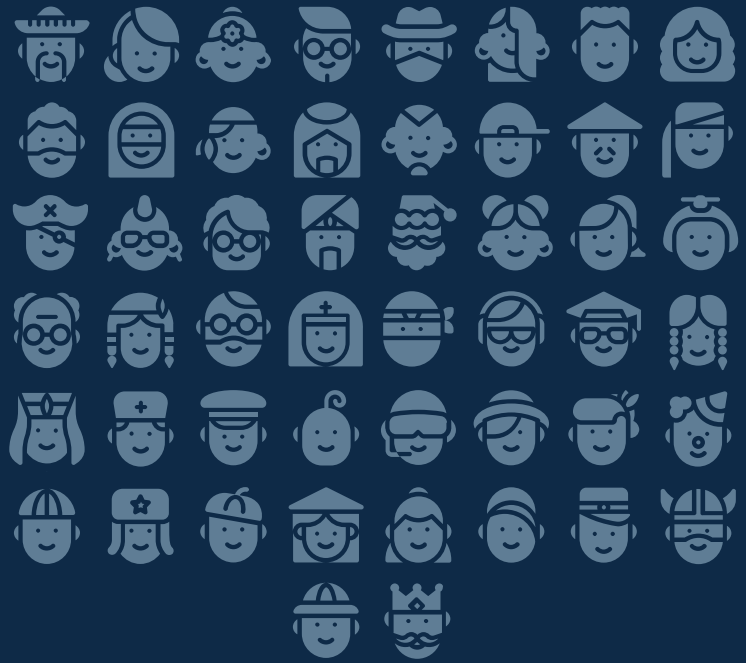
## Business Icons

## Teamwork Icons

## Help & Support Icons



## Avatar Icons



## Creative Process Icons


## Performing Arts Icons

# Nature Icons

# SEO & Marketing Icons


slidesgo