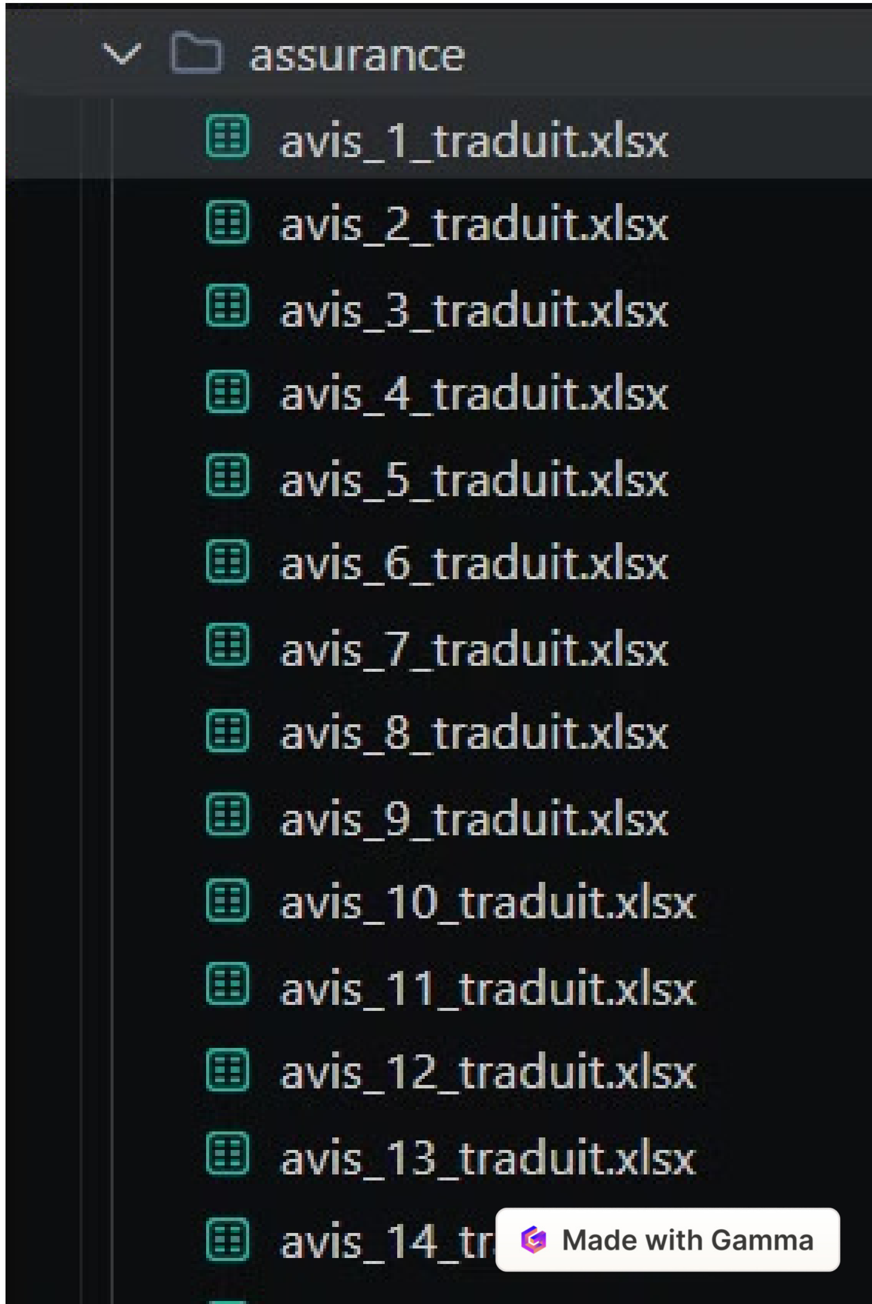
Project 2 - Insurance

Load Data

```
dfs = []
data_folder_path = os.path.join('assurance')
for file_name in os.listdir(data_folder_path):
    path = os.path.join(data_folder_path, file_name)
    new_df = pd.read_excel(path)
    dfs.append(new_df)
df = pd.concat(dfs, ignore_index=True)
```



```
assurance
    avis_1_traduit.xlsx
    avis_2_traduit.xlsx
    avis_3_traduit.xlsx
    avis_4_traduit.xlsx
    avis_5_traduit.xlsx
    avis_6_traduit.xlsx
    avis_7_traduit.xlsx
    avis_8_traduit.xlsx
    avis_9_traduit.xlsx
    avis_10_traduit.xlsx
    avis_11_traduit.xlsx
    avis_12_traduit.xlsx
    avis_13_traduit.xlsx
    avis_14_tr
```

Preprocess

```
# Select
drop_colmun = [
    'auteur', 'produit', 'date_publication', 'date_exp',
    'avis_cor', 'avis_cor_en']
df.drop(drop_colmun, axis=1, inplace=True)
df.dropna(subset=['avis_en'], inplace=True)
df['note'] = df['note'].fillna(0).astype(int)

# Rename columns
df.rename(columns={'avis': 'avis_fr', 'assureur': 'insurer'}, inplace=True)
df = df.reindex(columns=['insurer', 'avis_fr', 'avis_en','note', 'type']) # no agr inplace

# Split df
df_train = df[df['type'] == 'train'].drop(columns=['type'])        # 24 103 lines
df_test = df[df['type'] == 'test'].drop(columns=['type', 'note'])   # 10 330lines
```

Sentiment Analysis

Choose tabularisai/multilingual-sentiment-analysis because

- multilingual so it's the same model for the review in french and english
- return an integer score [0, 4] => translation to [1, 5] scale like stars

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification

senti_model_name = "tabularisai/multilingual-sentiment-analysis"
senti_tokenizer = AutoTokenizer.from_pretrained(senti_model_name)
senti_model = AutoModelForSequenceClassification.from_pretrained(
    senti_model_name).to(device)
senti_model = senti_model.eval()

def sentiment_pipeline(texts):
    inputs = senti_tokenizer(
        texts, return_tensors="pt", truncation=True, padding=True, max_length=512
    ).to(device)

    with torch.no_grad():
        outputs = senti_model(**inputs)
        probabilities = torch.nn.functional.softmax(outputs.logits, dim=-1)
        sentiment = torch.argmax(probabilities, dim=-1).tolist()

    senti_rescale = [int(senti+1) for senti in sentiment]    # convert to 1-5 scale like stars
    return senti_rescale
```

Zero-shot

Average distance with ground-truth

- Language fr : 0.79
- Language en : 0.79

Fine-tune on df_train

Didn't succeed

Subject Classifier

Choose cross-encoder/nli-deberta-v3-base because

- Finetune of the model microsoft/deberta-v3-base on SNLI and MultiNLI datasets, so it's also multilingual
- Return a score of probability devide between the different possible label

Pipeline

split_long_reviews

1. Split review in smaller chunks
2. Store the chunks with the other reviews
3. Keep a log of the original index of each review/subreview on id_map

merdge_splited_review

1. Merdge chunks with the same index in id_map
2. Compute average of each label if there is a merdge
3. Return the best label with it score
4. If the score is under a threshold, return 'Other'

```
from transformers import pipeline
from numpy import argmax

classifier = pipeline("zero-shot-classification",
    model='cross-encoder/nli-deberta-v3-base',
    use_fast = False, device=device
)

def subject_pipeline(reviews, lang='fr', threshold=0.5):
    # Pre-process
    labels = labels_fr if lang == 'fr' else labels_en
    split_reviews, id_map = split_long_reviews(reviews, 500)

    # Run model (exclude 'Others' label)
    resp = classifier(split_reviews, labels[:-1])

    # Post process
    subjects, scores = merdge_splited_review(resp, id_map)

    # Classify as 'Others' under thresholds
    for i in range(len(scores)):
        if scores[i] <= threshold:
            subjects[i] = labels[-1]

    return subjects
```

Summarize reviews

```
from transformers import pipeline

summarizer = pipeline("summarization",
    model="Falconsai/text_summarization",
    device=device
)

def summarize_reviews(summary_text, lang='fr', input_length_max=512,
output_length_max=300):
    nb_loop = 0
    print(f'Original length : {len(summary_text)} lines')
    print('Nb loop \t Summary length')

    # Summarize and merge until we have only one summary
    while len(summary_text) > 1:
        summary_merdge = join_reviews(summary_text, input_length_max)
        summary_resp = summarizer(summary_merdge,
            max_length=output_length_max,
            min_length=50,
            do_sample=False
        )
        summary_text = [r['summary_text'] for r in summary_resp]

    return summary_text[0]
```

`join_reviews`

Merge summaries till get "super review" of length just below `input_length_max`

If some reviews are already taller, apply a recursion on it

`summarizer`

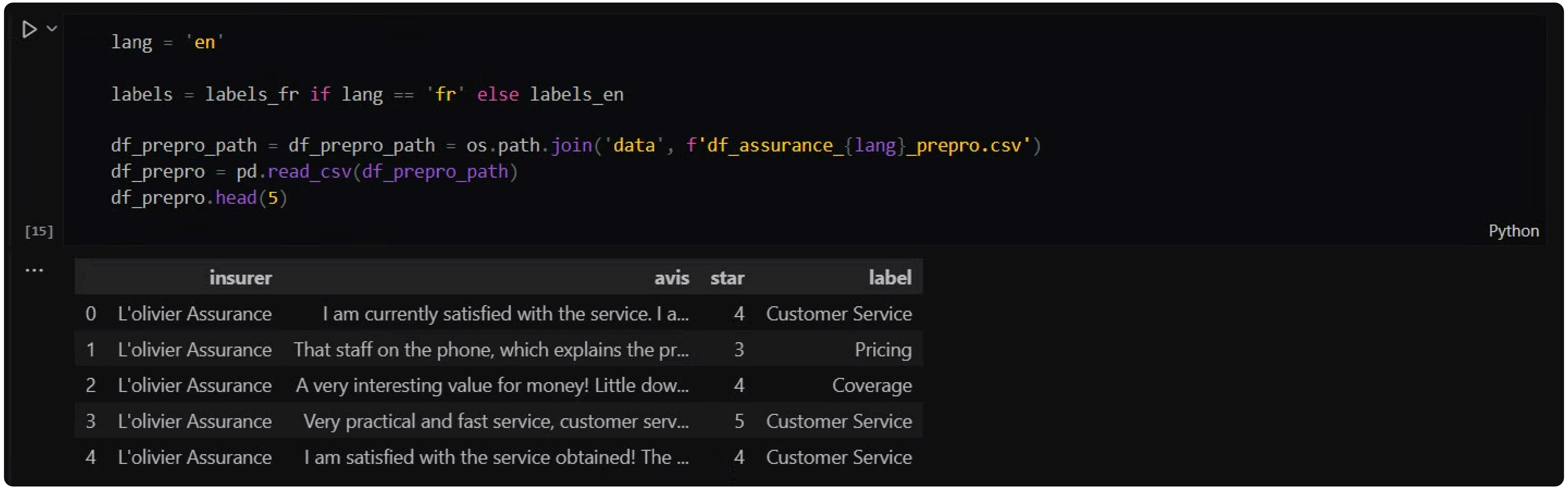
Summarize the reviews till length [`output_length_max`, 50]

Do it while we have one summary

Preprocess data test

For both language:

- Apply both sentiment_pipeline and subject_pipeline to the whole dataset
- Remove the reviews for the other language
- Save it as csv



```python
lang = 'en'

labels = labels_fr if lang == 'fr' else labels_en

df_prepro_path = df_prepro_path = os.path.join('data', f'df_assurance_{lang}_prepro.csv')
df_prepro = pd.read_csv(df_prepro_path)
df_prepro.head(5)
```

[15] Python

	insurer	avis	star	label
0	L'olivier Assurance	I am currently satisfied with the service. I a...	4	Customer Service
1	L'olivier Assurance	That staff on the phone, which explains the pr...	3	Pricing
2	L'olivier Assurance	A very interesting value for money! Little dow...	4	Coverage
3	L'olivier Assurance	Very practical and fast service, customer serv...	5	Customer Service
4	L'olivier Assurance	I am satisfied with the service obtained! The ...	4	Customer Service

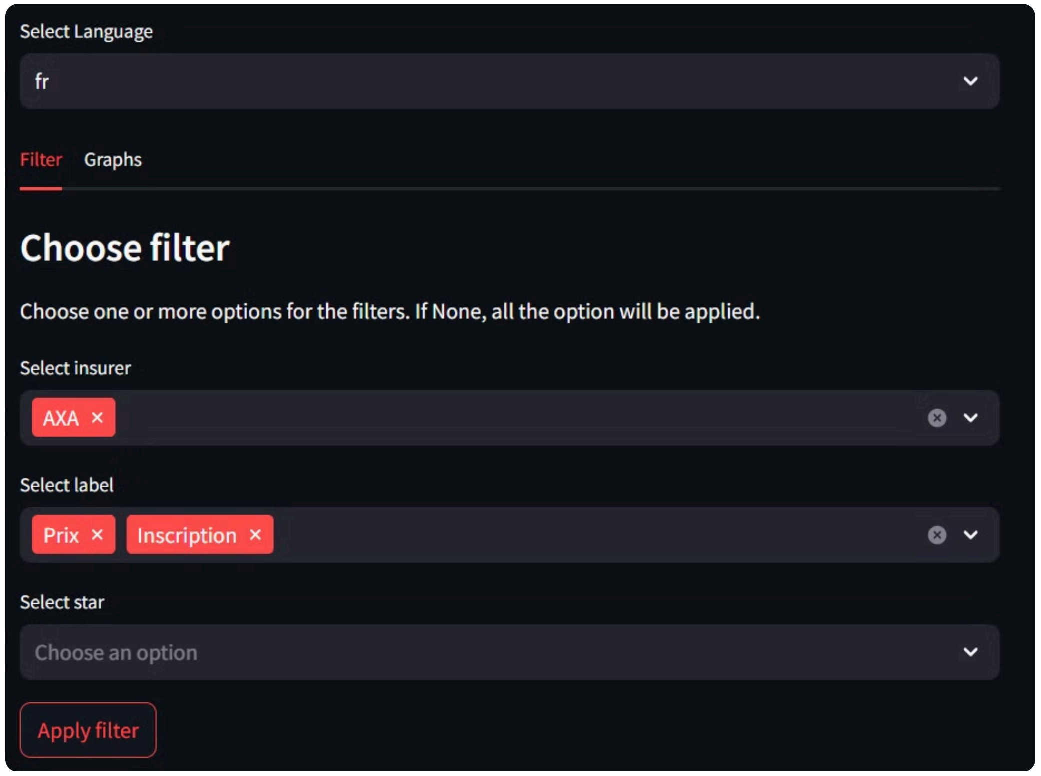
Made with Gamma

Streamlit Filter

```
def dataframe_filter(df_input, insurer=None, label=None, star=None):
    df_filter = df_input.copy()
    filter_dict = {'insurer': insurer, 'label': label, 'star': star }

    for filter_name, filter in filter_dict.items():
        if filter not in [None, []]:                    # if str or [str]
            filter = [filter] if type(filter) == str else filter    # convert str to [str]
            df_filter = df_filter[df_filter[filter_name].isin(filter)]

    return df_filter
```



Choose filter

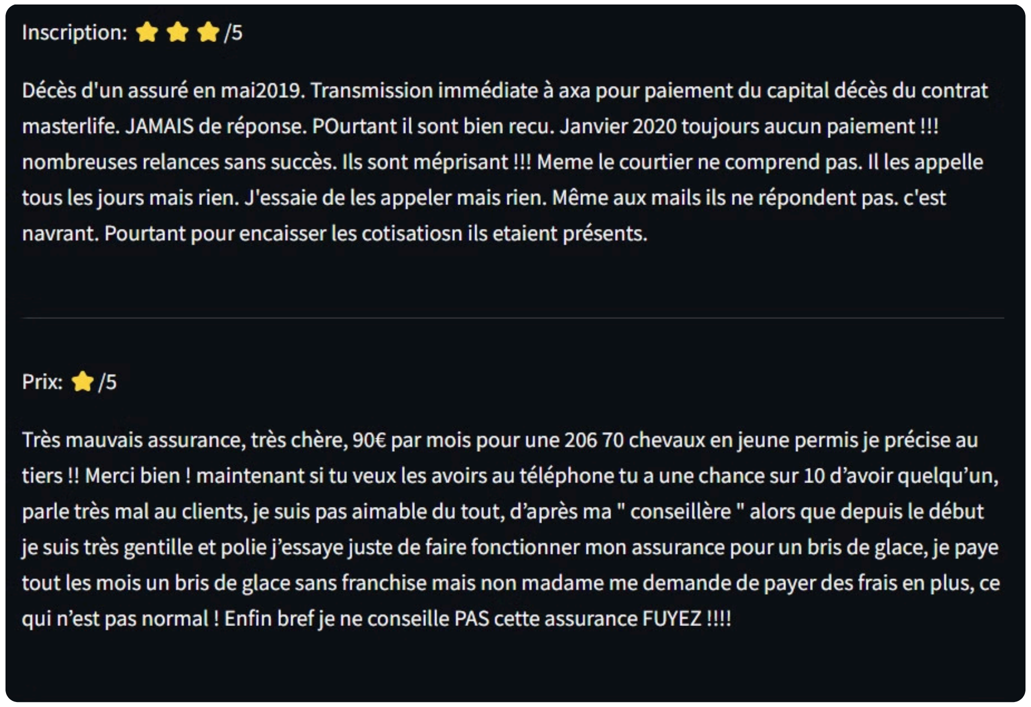
Choose one or more options for the filters. If None, all the option will be applied.

Select insurer: AXA

Select label: Prix, Inscription

Select star: Choose an option

Ask for the filter



Inscription: ⭐⭐⭐/5

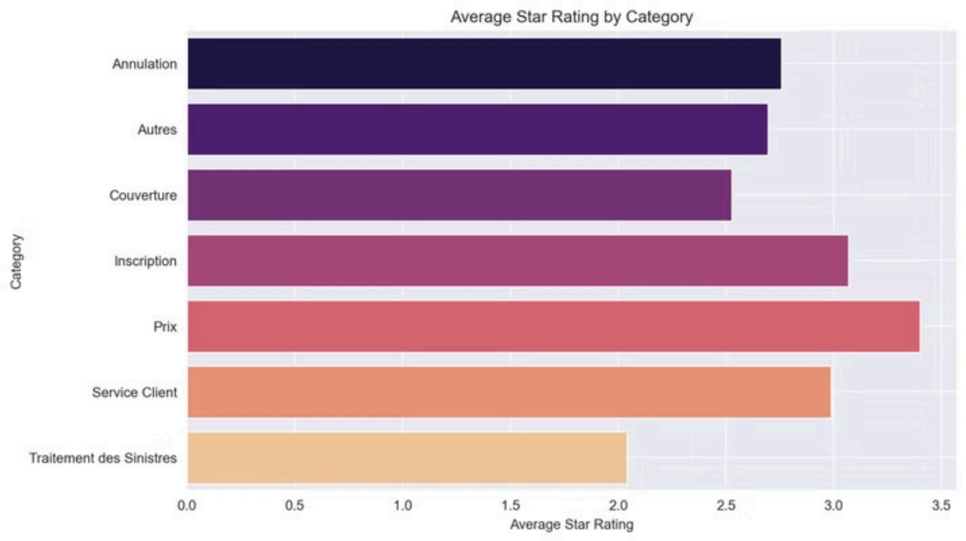
Décès d'un assuré en mai2019. Transmission immédiate à axa pour paiement du capital décès du contrat masterlife. JAMAIS de réponse. POurtant il sont bien recu. Janvier 2020 toujours aucun paiement !!! nombreuses relances sans succès. Ils sont méprisant !!! Meme le courtier ne comprend pas. Il les appelle tous les jours mais rien. J'essaie de les appeler mais rien. Même aux mails ils ne répondent pas. c'est navrant. Pourtant pour encaisser les cotisatiosn ils etaient présents.

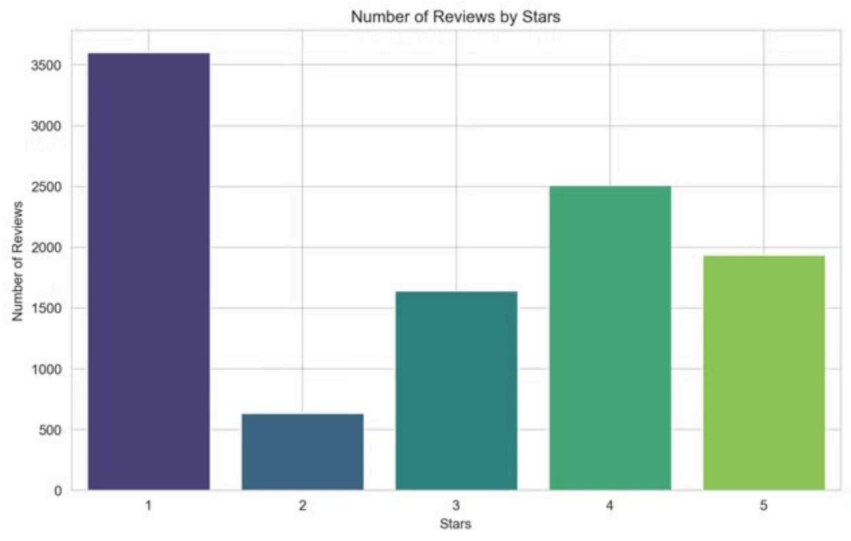
Prix: ⭐/5

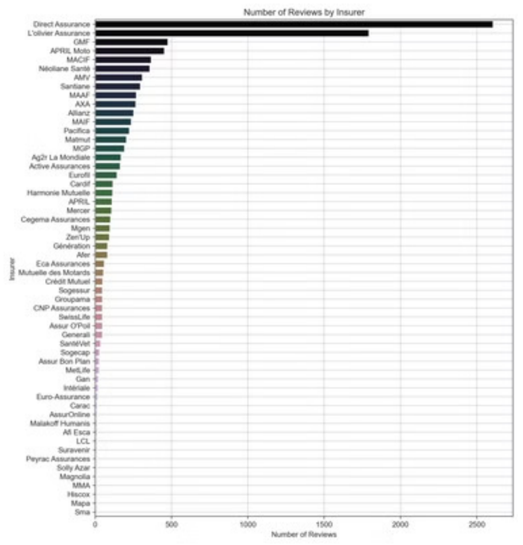
Très mauvais assurance, très chère, 90€ par mois pour une 206 70 chevaux en jeune permis je précise au tiers !! Merci bien ! maintenant si tu veux les avoirs au téléphone tu a une chance sur 10 d'avoir quelqu'un, parle très mal au clients, je suis pas aimable du tout, d'après ma " conseillère " alors que depuis le début je suis très gentille et polie j'essaye juste de faire fonctionner mon assurance pour un bris de glace, je paye tout les mois un bris de glace sans franchise mais non madame me demande de payer des frais en plus, ce qui n'est pas normal ! Enfin bref je ne conseille PAS cette assurance FUYEZ !!!!

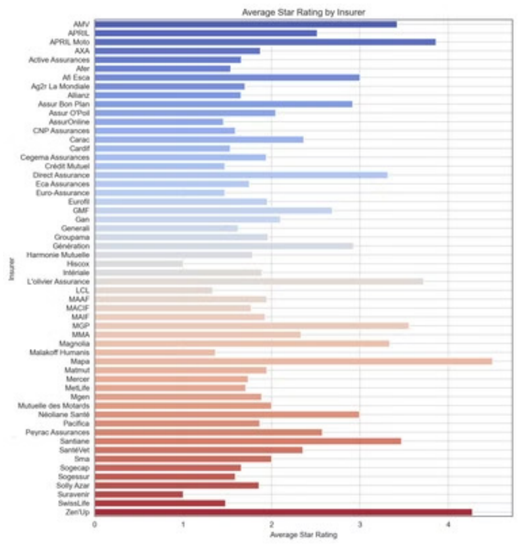
Print the result

Streamlit Graph

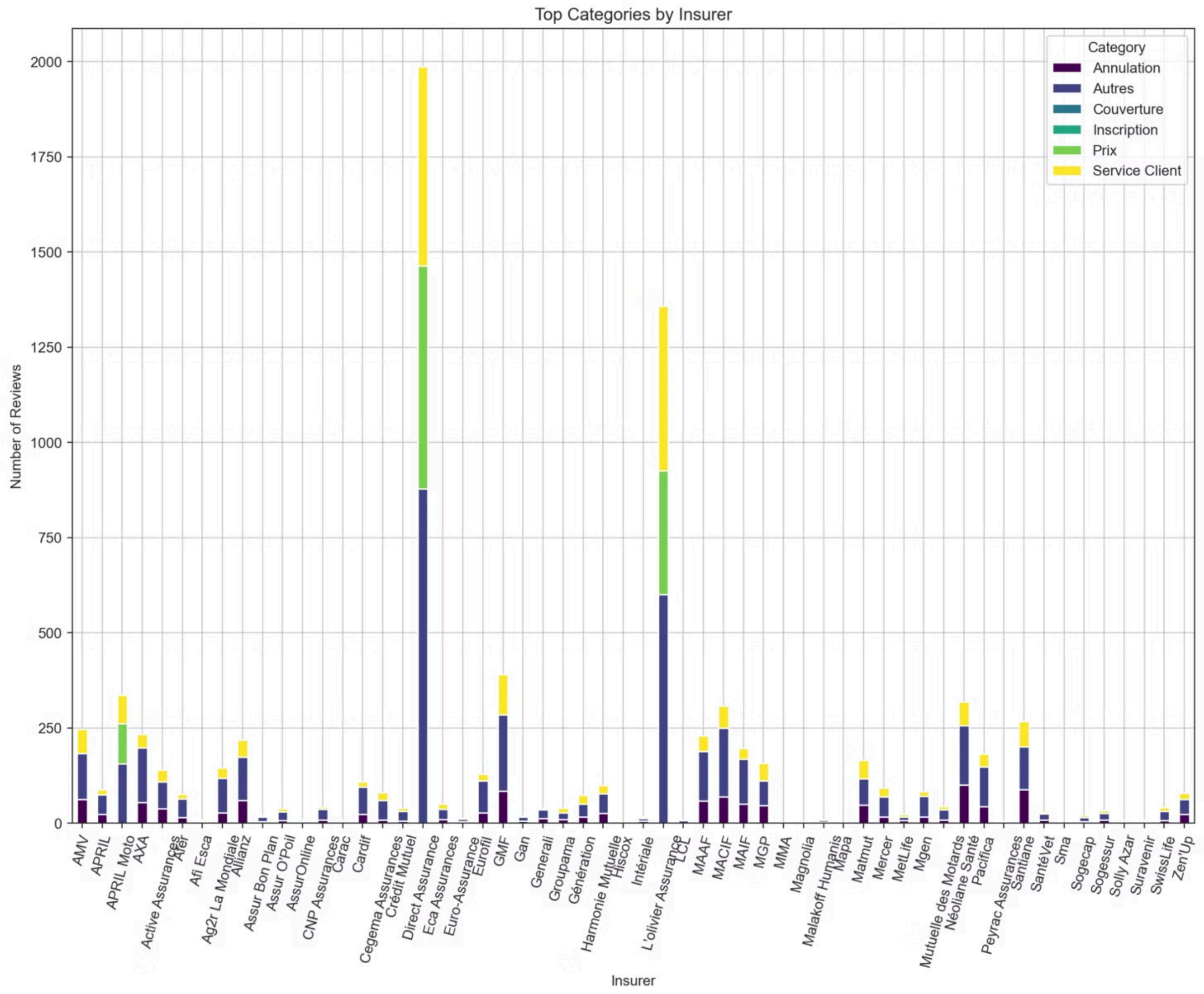








Made with Gamma


Top Categories by Insurer
Number of Reviews
0
250
500
750
1000
1250
1500
1750
2000
Category
Annulation
Autres
Couverture
Inscription
Prix
Service Client
AMV
APRIL
APRIL Moto
AXA
Active Assurances
Afer
Afi Esca
Ag2r La Mondiale
Allianz
Assur Bon Plan
Assur O'Poil
AssurOnline
CNP Assurances
Carac
Cardif
Cegema Assurances
Crédit Mutuel
Direct Assurance
Eca Assurances
Euro-Assurance
Eurofil
GMF
Gan
Generali
Groupama
Génération
Harmonie Mutuelle
Hiscox
Intériale
L'olivier Assurance
LCL
MAAF
MACIF
MAIF
MGP
MMA
Magnolia
Malakoff Humanis
Mapa
Matmut
Mercer
MetLife
Mgen
Mutuelle des Motards
Néoliane Santé
Pacifica
Peyrac Assurances
Santiane
SantéVet
Sma
Sogecap
Sogessur
Solly Azar
Suravenir
SwissLife
Zen'Up
Insurer