Category 1

It has been seen time and time again that minorities in the US and elsewhere, In the US these include people of colour/African Americans, etc are unfairly targeted in the following areas.

This is not a comprehensive list.

1. **Loans**

   Banks/credit card companies, etc use machine learning models which predict how likely would-be borrower is likely to miss loan payments. A majority of people who take loans do pay them back on time. Hence the datasets are likely to have more cases where folk

   s paid loans on time than those who didn't, i.e unbalanced. If you look at the part of the dataset which has data on loans which had missed payments/defaulted, it is likely to show that most people who fall in this category are minorities.

   This has societal ramifications. If as a person of colour you apply for a loan/credit card, you could be charged higher interest rates which make the monthly payments high, or you could be denied loans altogether. For families in the poorer section of society, this might mean children being unable to go to college, unable to purchase homes/credit cards etc.

2. **Auto insurance premium**

   Due to racial segregation and gentrification in recent times, neighbourhoods have become homogenous. Neighbourhoods with the majority of 8whites/Asian/minorities. As a person of colour, you are more likely to live in a neighbourhood which is the majority of people of colour. If you apply for auto insurance, you are likely to be charged higher premiums. Drivers who live in these communities are on average quoted premiums that are 70 per cent more expensive than premiums for drivers who live in white communities—an average of $438 dollars per year more. Minorities paying more for auto insurance

3. **Crimes**

   As a person of colour, you are likely to live in a neighbourhood with other people of colour. These people are likely to belong to a lower socioeconomic status. Part-time jobs/single parent household/lots of children in these homes/unable to send kids to school, etc.

   If the crime datasets used by law enforcement is biased, i.e. most people likely to commit crime are people of colour, it is very likely that as a person of colour you

are likely to be stopped by the highway patrol, unnecessarily arrested/longer sentences for the same crime, etc.
https://openpolicing.stanford.edu/

Datasets of possible use:
https://www.huduser.gov/portal/datasets/hsg_discrimination.html

Problem Statement:

From the past studies and analysis, it has been observed that the areas that come under minority regions are unfairly targeted by auto insurance companies by keeping the premiums higher than average in the minority. They argue that the premiums are high due to the fact that these areas are at greater risk of accidents. The companies also claim that their models are colour-blind and completely based on these risks. In our project, we will be examining this claim and find out whether or not the claim made by insurance companies is actually true. While analyzing the data, we will build a model which predicts the premium of the auto-insurance paid by the people living in different neighbourhoods. Based on the model, we will see if there is a bias present. If true, we will come to the conclusion that the people living in minority areas are unfairly targeted by insurance companies.

Finding Racial Discrimination in Home Loan Approvals

We will be dealing with two different datasets here.

HMDA : https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=il&records=all-records&field_descriptions=labels

We use this data to get the information about the loan applications from a few different states.

Census Data : https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

This dataset will be used to find the population proportions with respect to the race. Different states have different proportions of populations present. So it may be the case that a particular state has less number of people in minority, leading to the fact that the loans approved here will be less for minorities, as the number of applications coming from minorities is less.

Steps involved in the analysis

1. Exploratory analysis - Initial investigations, finding patterns in the dataset and framing hypothesis. Check for the assumptions that we make in the initial stage.
2. In the initial stage, our focus will be on one state. We will do all the analysis on the dataset of one state only. Once we are done with it, we will consider other states as the project progresses and integrate each module at the end. The primary task will be to frame a hypothesis and come to the inference.

Category 2 (Health insurance fraudulent claims)
https://www.nhcaa.org/resources/health-care-anti-fraud-resources/the-challenge-of-health-care-fraud.aspx
While a majority of health care claims in the US are legitimate, there is a small percentage which is fraudulent. This is in the form of unnecessary expensive tests conducted on patients, reporting tests when they weren't performed etc.
Would we be able to build a machine learning model which could predict the likelihood of a claim being fraudulent? This would require a dataset, understanding which providers are likely to commit such frauds, etc.

NLP
Areas include text summarization/named entity recognition/parts of speech detection/machine translation/lyrics, story generation, etc. The last three are an application of recurrent neural networks (LSTM) or Markov chains.

Other categories? (Please list them here)
Computer vision/image processing. Could intersect with deep learning-based models. This is a broad area. Needs problem identification before figuring out datasets. Possible areas include image classification/localization, image segmentation, semantic segmentation, etc.

Category 3

Instagram Hashtag Generation
Predict Instagram hashtag according to the picture given as an input