**Week 7 Day 4**

**Ex1 Identify Data Types**

Below are various data sources. Identify whether each one is an example of structured or unstructured data.

* A company’s financial reports stored in an Excel file.
  + Structured
* Photographs uploaded to a social media platform.
  + Unstructured
* A collection of news articles on a website.
  + Unstructured
* Inventory data in a relational database.
  + Structured
* Recorded interviews from a market research study.
  + Unstructured

**Ex 2 Transformation Exercise**

For each of the following unstructured data sources, propose a method to convert it into structured data. Explain your reasoning.

* A series of blog posts about travel experiences.
  + In the structured dB you can have a table for countries, cities, users, and blogs. So, in the countries, cities, and users’ tables you will have many more data points that will link to the blogs table via an id number. This will be beneficial from a Data Analyst point because you will have much more information other than just the text of the blog post – information of where the travel was experienced e.g. country GDP, city density population, country religion.
* Audio recordings of customer service calls.
  + In the structured dB you can have a table for the customer, the employee, and the call information. By doing this it will be easier to see the demographics of all customers, and give more insight into possible reasons for customers that call in greater numbers
* Handwritten notes from a brainstorming session.
  + In the structured dB you can have a fixed table for each session which includes an upload of the handwritten notes. By having a structured dB you can use multiple linked tables to analyse the trends in brainstorming sessions – and why they are occurring more for some categories and less for others.
* A video tutorial on cooking.
  + In the structured dB you can have one table for the video – with the title, subtitle, duration. Then you can have other tables for the ingredients of the food being cooked, a table for the step-by-step written guide (each step is a new row). Then you can also have a table for the comments of the video, and also feedback. By having these structured tables it will be easier to see where the video can be improved, and also allows to give more information to the customer.

**Ex 3 Application Scenario**

You are a data analyst at a retail company. You have access to various data sources, including transaction records, customer feedback comments, social media posts about your brand, and employee work schedules.

* Categorize each of these data sources as structured or unstructured.
  + transaction records - structured
  + customer feedback comments - unstructured
  + social media posts about your brand - unstructured
  + employee work schedules – structured
* Suggest how you might use each type of data for improving the company’s business operations.
  + transaction records
    - I would you this data to see the trends of the transactions, the products sold, the days and times of days when sales are at their highest. From this data we will be able to push sales for products that aren’t selling well (advertising, discounts), and work on a marketing strategy to improve the customer intake on days and or times of the day when the business is quiet.
  + customer feedback comments
    - I would use this data to understand the customers more than just their transactions. Where our company can improve for them, and why we need to improve certain aspects of the products or company itself.
  + social media posts about your brand
    - I would use this data to understand the demographics (age, location) of where the brand is doing well and getting positive feedback and where it is getting negative feedback – so we can improve on those areas.
  + employee work schedules
    - I would use this data to see how many days/hours a week all employees are working, and how many days off each employee has. From this analysis I will be able to see if the employees are overworked and we need to start hiring more people. Or if there are opportunities for employees to work more, but in a better way.

**Ex 4 Synthetic Data Generation**

In ex4-synthetic\_data\_generation file

**Ex 5 Data Augmentation for Images**

**Ex6 Simulation-based dataset creation**

* Imagine you are developing a simulation for traffic flow in a city. Your goal is to generate a dataset that reflects different traffic conditions at various times of the day.
* *Outline the steps you would take to create this simulation. Consider factors like vehicle types, road types, traffic signals, and peak/off-peak hours.*
  + I would create lists of different vehicle types, road types, and the different levels of congestion during certain hours
  + With these lists it will help to understand the capacity (number of cars) a specific road type can handle during the hours in the list
  + Then with the understanding of the capacity of each road type we can then run a simulation of a traffic flow with different types of roads at different hours with different levels of capacity and different vehicle types to see where the flow may work and where it won’t work
* *Describe how you would collect data from this simulation, specifying the types of data (e.g., vehicle count, average speed) you would gather at different time intervals.*
  + The collection of the data from the simulation will be broken down into different road types and then broken down further to every hour of the day. Within each hour we will look at the count of all vehicles for each hour, the count of each distinct type of vehicle, the average speed of all vehicles, the minimum speed of all vehicles, the maximum speed of all vehicles, and the location of the minimum speeds.
* *Discuss the potential uses of this simulated dataset in traffic management and urban planning.*
  + We can use the data to understand where blockages occur in different roads, and how a traffic flow can be improved to get a better average speed and lower count of vehicles per hour – thus ensuring the flow is continuous and doesn’t stagnate at certain hours of the day