|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set | **Description** | **Potential applications of Machine Learning** | **Related Industry** | **Impact** | **Value proposition** |
| US accidents | This is a countrywide car accident dataset that covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records. | -Predicting accident severity and likelihood by some environmental and location features. | Transportation, Insurance, infrastructure | Reduced accidents and traffic congestion  - Improved safety measures  - Enhanced insurance risk assessment  - Efficient urban development | By the final model, the likelihood of accident and its severity can be predicted. |
| Cancer data | These data were aggregated from a number of sources including the American Community Survey ([census.gov](http://census.gov/)), [clinicaltrials.gov](http://clinicaltrials.gov/), and [cancer.gov](http://cancer.gov/). This data set contains variables about death rate and some social and medical situation of patients for 2010 – 2016 period. | -classification problem of cancer patient based on some economic and demographic features | Healthcare, Research, sociology | -- Enhanced understanding of cancer socio-economic causes  - Informed healthcare policies | improve cancer care |
| Road accident data | This comprehensive dataset provides detailed information on road accidents reported over multiple years. The dataset encompasses various attributes related to accident status, vehicle and casualty references, demographics, and severity of casualties. It includes essential factors such as pedestrian details, casualty types, road maintenance worker involvement, and the Index of Multiple Deprivation (IMD) decile for casualties' home areas. | - Accident causality analysis and prevention strategies  - Road safety prediction and improvement recommendations  - Targeted public safety interventions  - Infrastructure planning and maintenance optimization | Transportation, Public Safety, Urban Planning | - Safer roads and reduced accidents  - Informed infrastructure planning  - Effective public safety policies  - Insights into accident causes | - Enhancing road safety |
| Ecommerce behavior data | This file contain behavior data for 7 months (from October 2019 to April 2020) from a large multi-category online store.  Each row in the file represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.  Data collected by [Open CDP](https://rees46.com/en/open-cdp) project. Feel free to use open source customer data platform. | - Customer segmentation and personalized product recommendations  - Fraud detection and prevention  - Customer churn prediction and retention strategies  - Sales forecasting and inventory optimization | Ecommerce, Marketing | - Improved customer experience and engagement  - Reduced fraud and increased security  - Higher customer retention and revenue  - Efficient inventory management | -improving market knowledge |

<https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents>

This is a countrywide car accident dataset that covers **49 states of the USA**. The accident data were collected from **February 2016 to March 2023**, using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately **7.7 million** accident records. For more information about this dataset, please visit [here](https://smoosavi.org/datasets/us_accidents).

Format

The data is provided in terms of a CSV file. Following table describes the data attributes (visit our paper to learn more about these attributes and how we obtained them):

# Attribute Description Nullable

1 ID This is a unique identifier of the accident record. No

2 Severity Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay). No

3 Start\_Time Shows start time of the accident in local time zone. No

4 End\_Time Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed. No

5 Start\_Lat Shows latitude in GPS coordinate of the start point. No

6 Start\_Lng Shows longitude in GPS coordinate of the start point. No

7 End\_Lat Shows latitude in GPS coordinate of the end point. Yes

8 End\_Lng Shows longitude in GPS coordinate of the end point. Yes

9 Distance(mi) The length of the road extent affected by the accident. No

10 Description Shows natural language description of the accident. No

11 Number Shows the street number in address field. Yes

12 Street Shows the street name in address field. Yes

13 Side Shows the relative side of the street (Right/Left) in address field. Yes

14 City Shows the city in address field. Yes

15 County Shows the county in address field. Yes

16 State Shows the state in address field. Yes

17 Zipcode Shows the zipcode in address field. Yes

18 Country Shows the country in address field. Yes

19 Timezone Shows timezone based on the location of the accident (eastern, central, etc.). Yes

20 Airport\_Code Denotes an airport-based weather station which is the closest one to location of the accident. Yes

21 Weather\_Timestamp Shows the time-stamp of weather observation record (in local time). Yes

22 Temperature(F) Shows the temperature (in Fahrenheit). Yes

23 Wind\_Chill(F) Shows the wind chill (in Fahrenheit). Yes

24 Humidity(%) Shows the humidity (in percentage). Yes

25 Pressure(in) Shows the air pressure (in inches). Yes

26 Visibility(mi) Shows visibility (in miles). Yes

27 Wind\_Direction Shows wind direction. Yes

28 Wind\_Speed(mph) Shows wind speed (in miles per hour). Yes

29 Precipitation(in) Shows precipitation amount in inches, if there is any. Yes

30 Weather\_Condition Shows the weather condition (rain, snow, thunderstorm, fog, etc.) Yes

31 Amenity A POI annotation which indicates presence of amenity in a nearby location. No

32 Bump A POI annotation which indicates presence of speed bump or hump in a nearby location. No

33 Crossing A POI annotation which indicates presence of crossing in a nearby location. No

34 Give\_Way A POI annotation which indicates presence of give\_way in a nearby location. No

35 Junction A POI annotation which indicates presence of junction in a nearby location. No

36 No\_Exit A POI annotation which indicates presence of no\_exit in a nearby location. No

37 Railway A POI annotation which indicates presence of railway in a nearby location. No

38 Roundabout A POI annotation which indicates presence of roundabout in a nearby location. No

39 Station A POI annotation which indicates presence of station in a nearby location. No

40 Stop A POI annotation which indicates presence of stop in a nearby location. No

41 Traffic\_Calming A POI annotation which indicates presence of traffic\_calming in a nearby location. No

42 Traffic\_Signal A POI annotation which indicates presence of traffic\_signal in a nearby loction. No

43 Turning\_Loop A POI annotation which indicates presence of turning\_loop in a nearby location. No

44 Sunrise\_Sunset Shows the period of day (i.e. day or night) based on sunrise/sunset. Yes

45 Civil\_Twilight Shows the period of day (i.e. day or night) based on civil twilight. Yes

46 Nautical\_Twilight Shows the period of day (i.e. day or night) based on nautical twilight. Yes

47 Astronomical\_Twilight Shows the period of day (i.e. day or night) based on astronomical twilight. Yes

Applications of Dataset

US-Accidents can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence.

<https://data.world/nrippner/ols-regression-challenge>

These data were aggregated from a number of sources including the American Community Survey ([census.gov](http://census.gov/)), [clinicaltrials.gov](http://clinicaltrials.gov/), and [cancer.gov](http://cancer.gov/). Most of the data preparation process can be veiwed [here](https://data.world/nrippner/cancer-trials).

**Data Dictionary**

**TARGET\_deathRate:** Dependent variable. Mean *per capita* (100,000) cancer mortalities(*a*)

**avgAnnCount:** Mean number of reported cases of cancer diagnosed annually(*a*)

**avgDeathsPerYear:** Mean number of reported mortalities due to cancer(*a*)

**incidenceRate:** Mean *per capita* (100,000) cancer diagoses(*a*)

**medianIncome:** Median income per county (*b*)

**popEst2015:** Population of county (*b*)

**povertyPercent:** Percent of populace in poverty (*b*)

**studyPerCap:** *Per capita* number of cancer-related clinical trials per county (*a*)

**binnedInc:** Median income per capita binned by decile (*b*)

**MedianAge:** Median age of county residents (*b*)

**MedianAgeMale:** Median age of male county residents (*b*)

**MedianAgeFemale:** Median age of female county residents (*b*)

**Geography:** County name (*b*)

**AvgHouseholdSize:** Mean household size of county (*b*)

**PercentMarried:** Percent of county residents who are married (*b*)

**PctNoHS18\_24:** Percent of county residents ages 18-24 highest education attained: less than high school (*b*)

**PctHS18\_24:** Percent of county residents ages 18-24 highest education attained: high school diploma (*b*)

**PctSomeCol18\_24:** Percent of county residents ages 18-24 highest education attained: some college (*b*)

**PctBachDeg18\_24:** Percent of county residents ages 18-24 highest education attained: bachelor's degree (*b*)

**PctHS25\_Over:** Percent of county residents ages 25 and over highest education attained: high school diploma (*b*)

**PctBachDeg25\_Over:** Percent of county residents ages 25 and over highest education attained: bachelor's degree (*b*)

**PctEmployed16\_Over:** Percent of county residents ages 16 and over employed (*b*)

**PctUnemployed16\_Over:** Percent of county residents ages 16 and over unemployed (*b*)

**PctPrivateCoverage:** Percent of county residents with private health coverage (*b*)

**PctPrivateCoverageAlone:** Percent of county residents with private health coverage alone (no public assistance) (*b*)

**PctEmpPrivCoverage:** Percent of county residents with employee-provided private health coverage (*b*)

**PctPublicCoverage:** Percent of county residents with government-provided health coverage (*b*)

**PctPubliceCoverageAlone:** Percent of county residents with government-provided health coverage alone (*b*)

**PctWhite:** Percent of county residents who identify as White (*b*)

**PctBlack:** Percent of county residents who identify as Black (*b*)

**PctAsian:** Percent of county residents who identify as Asian (*b*)

**PctOtherRace:** Percent of county residents who identify in a category which is not White, Black, or Asian (*b*)

**PctMarriedHouseholds:** Percent of married households (*b*)

**BirthRate:** Number of live births relative to number of women in county (*b*)

(*a*): years 2010-2016

(*b*): 2013 Census Estimates

<https://www.kaggle.com/datasets/juhibhojani/road-accidents-data-2022>

Description:  
This comprehensive dataset provides detailed information on road accidents reported over multiple years. The dataset encompasses various attributes related to accident status, vehicle and casualty references, demographics, and severity of casualties. It includes essential factors such as pedestrian details, casualty types, road maintenance worker involvement, and the Index of Multiple Deprivation (IMD) decile for casualties' home areas.

1. **Status**: The status of the accident (e.g., reported, under investigation).
2. **Accident\_Index**: A unique identifier for each reported accident.
3. **Accident\_Year**: The year in which the accident occurred.
4. **Accident\_Reference**: A reference number associated with the accident.
5. **Vehicle\_Reference**: A reference number for the involved vehicle in the accident.
6. **Casualty\_Reference**: A reference number for the casualty involved in the accident.
7. **Casualty\_Class**: Indicates the class of the casualty (e.g., driver, passenger, pedestrian).
8. **Sex\_of\_Casualty**: The gender of the casualty (male or female).
9. **Age\_of\_Casualty**: The age of the casualty.
10. **Age\_Band\_of\_Casualty**: Age group to which the casualty belongs (e.g., 0-5, 6-10, 11-15).
11. **Casualty\_Severity**: The severity of the casualty's injuries (e.g., fatal, serious, slight).
12. **Pedestrian\_Location**: The location of the pedestrian at the time of the accident.
13. **Pedestrian\_Movement**: The movement of the pedestrian during the accident.
14. **Car\_Passenger**: Indicates whether the casualty was a car passenger at the time of the accident (yes or no).
15. **Bus\_or\_Coach\_Passenger**: Indicates whether the casualty was a bus or coach passenger (yes or no).
16. **Pedestrian\_Road\_Maintenance\_Worker**: Indicates whether the casualty was a road maintenance worker (yes or no).
17. **Casualty\_Type**: The type of casualty (e.g., driver/rider, passenger, pedestrian).
18. **Casualty\_Home\_Area\_Type**: The type of area in which the casualty resides (e.g., urban, rural).
19. **Casualty\_IMD\_Decile**: The IMD decile of the area where the casualty resides (a measure of deprivation).
20. **LSOA\_of\_Casualty**: The Lower Layer Super Output Area (LSOA) associated with the casualty's location.

This dataset provides valuable insights for analyzing road accidents, identifying trends, and implementing safety measures to reduce casualties and enhance road safety. Researchers, policymakers, and analysts can leverage this dataset for evidence-based decision-making and improving overall road transportation systems.

* <https://www.kaggle.com/datasets/mkechinov/ecommerce-behavior-data-from-multi-category-store>

| **Property** | **Description** |
| --- | --- |
| **event\_time** | Time when event happened at (in UTC). |
| **event\_type** | Only one kind of event: purchase. |
| **product\_id** | ID of a product |
| **category\_id** | Product's category ID |
| **category\_code** | Product's category taxonomy (code name) if it was possible to make it. Usually present for meaningful categories and skipped for different kinds of accessories. |
| **brand** | Downcased string of brand name. Can be missed. |
| **price** | Float price of a product. Present. |
| **user\_id** | Permanent user ID. |
| \*\* user\_session\*\* | Temporary user's session ID. Same for each user's session. Is changed every time user come back to online store from a long pause. |

**Event types**

Events can be:

* view - a user viewed a product
* cart - a user added a product to shopping cart
* remove\_from\_cart - a user removed a product from shopping cart
* purchase - a user purchased a product

**Multiple purchases per session**

A session can have multiple **purchase** events. It's ok, because it's a single order.