Data Cleaning

AUTHOR
Nandini Kodali

Introduction

Raw data collected from various sources is rarely in a format ready for analysis. It often contains inconsistencies, missing values, duplicates, and irrelevant information, which can hinder the analytical process and lead to inaccurate or biased results. **Data Cleaning** is the process of transforming this messy data into a structured, consistent, and reliable format, suitable for extracting meaningful insights and applying models effectively.

Importance:

- Data cleaning ensures that the data is complete, accurate, and consistent, significantly enhancing the reliability of insights derived from the analysis and the performance of models built using the data. High-quality data forms the foundation of robust models and predictions.
- Missing values in the dataset can introduce bias or distort the analysis. Addressing these gaps through techniques like **imputation** (filling in missing values using statistical or logical methods) or **removal** helps maintain the integrity of results.
- Duplicates in the dataset can overrepresent certain patterns, while outliers may distort metrics and affect model accuracy. Identifying and appropriately handling these anomalies ensures the analysis remains valid and unbiased.
- When data is collected from multiple sources, differences in format, naming conventions, and measurement units can cause inconsistencies. Standardizing these elements across all datasets ensures that the data can be seamlessly integrated and analyzed as a whole.

In this project:
Data was collected from a variety of sources, Data Cleaning was an essential step. Each raw dataset underwent a tailored cleaning process designed to fit its specific structure and use case. These processes will be discussed in detail within the respective sections dedicates to each dataset.

Data Cleaning

Provide the source code used for this section of the project here.

If you're using a package for code organization, you can import it at this point. However, make sure that the **actual workflow steps**—including data processing, analysis, and other key tasks—are conducted and clearly demonstrated on this page. The goal is to show the technical flow of your project, highlighting how the code is executed to achieve your results.

If relevant, link to additional documentation or external references that explain any complex components. This section should give readers a clear view of how the project is implemented from a technical perspective.

Remember, this page is a technical narrative, NOT just a notebook with a collection of code cells, include in-line Prose, to describe what is going on.

Required Libraries

▶ Code

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/nandinikodali/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

News Data

▶ Code

▶ Code

▶ Code

▶ Code

▶ Code

```
RAW DATA---
Norris calls Verstappen 'dangerous' as Sainz wins in Mexico : Lando
Norris cuts Max Verstappen’s lead to 47 points and labels his rival
“dangerous” as the championship battle reaches boiling point at the
Mexico City Grand Prix.
How well do you know Fernando Alonso? : As he prepares for his
400th F1 grand prix in Mexico City this weekend, find out how much
you know about Fernando Alonso.


CLEAN DATA---
norris calls verstappen 'dangerous' sainz wins mexico : lando
norris cuts verstappens lead points labels rival dangerous
championship battle reaches boiling point mexico city grand prix
well know fernando alonso : prepares grand prix mexico city weekend
find much know fernando alonso
```

Drivers Standings

▶ Code

▶ Code

▶ Code

	Season	Position	FirstName	LastName	Constructor_ID	Constructor_Name	Po
0	2000	1	Michael	Schumacher	ferrari	Ferrari	10
1	2000	2	Mika	Häkkinen	mclaren	McLaren	89
2	2000	3	David	Coulthard	mclaren	McLaren	73
3	2000	4	Rubens	Barrichello	ferrari	Ferrari	62
4	2000	5	Ralf	Schumacher	williams	Williams	24

▶ Code

```
Season              0
Position            0
FirstName           0
LastName            0
Constructor_ID      0
Constructor_Name    0
Points              0
Wins                0
dtype: int64
```

▶ Code

Circuit Information

▶ Code

▶ Code

▶ Code

	Circuit_ID	Circuit_Name	Country	Latitude	Longitude
0	adelaide	Adelaide Street Circuit	Australia	-34.9272	138.617
1	ain-diab	Ain Diab	Morocco	33.5786	-7.6875
2	aintree	Aintree	UK	53.4769	-2.94056
3	albert_park	Albert Park Grand Prix Circuit	Australia	-37.8497	144.968
4	americas	Circuit of the Americas	USA	30.1328	-97.6411

▶ Code

```
Circuit_ID      0
Circuit_Name    0
Country         0
Latitude        0
Longitude       0
dtype: int64
```

▶ Code

Race data

▶ Code

▶ Code

	season	round	raceName	url	circuit
0	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
1	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
2	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
3	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
4	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit

5 rows × 28 columns

▶ Code

Weather Data

▶ Code

▶ Code

```
season      0
raceName    0
url         0
weather     0
dtype: int64
```

▶ Code

```
0                                                Sunny
1                      Overcast with light rain at start
2                                  Mainly cloudy, dry
3                                        Cloudy, rain
```

```
4                                    Mainly cloudy, dry
                                ...
117    Sunny with temperatures reaching up to 27 °C (...
118    Dry start, with heavy rain and thunderstorm/mo...
119                                                 Rain
120                                                Sunny
121                                          Warm, Sunny
Name: weather, Length: 122, dtype: object
```

we will try to categorise the weather description into one of the following categories:

1. Sunny
2. Cloudy
3. Rainy
4. Windy

▶ Code

▶ Code

▶ Code

```
weather_class
Sunny            95
Cloudy           24
Rainy             2
Not Available     1
Name: count, dtype: int64
```

Due to severe class imbalance, `weather_class` feature is not a suitable candidate for effective analysis or model training.

▶ Code

```
   season             raceName  \
9    2006  European Grand Prix

                                                  url        weather
\
9  http://en.wikipedia.org/wiki/2006_European_Gra...  Not Available

   weather_class
9  Not Available
```

The weather data for 2006 European Grand Prix is not available on wikipedia.

- Using longitude, latitude and the date: The weather was `Sunny`

▶ Code

▶ Code

Race Info Merged

▶ Code

▶ Code

	season	round	raceName	url	circuit
0	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
1	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
2	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
3	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
4	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit

5 rows × 29 columns

▶ Code

```
Index(['season', 'round', 'raceName', 'url', 'circuitName',
'locality',
       'country', 'lat', 'long', 'date', 'number', 'position',
'positionText',
       'points', 'Driver', 'Constructor', 'grid', 'laps', 'status',
'Time',
       'FastestLap', 'driverId', 'driverGivenName',
'driverFamilyName',
       'constructorId', 'constructorName', 'timeMillis', 'time',
       'weather_class'],
      dtype='object')
```

▶ Code

▶ Code

```
Index(['season', 'round', 'raceName', 'url', 'circuitName',
'locality',
       'country', 'date', 'position', 'points', 'grid', 'laps',
'status',
       'driverId', 'driverGivenName', 'driverFamilyName',
'constructorId',
       'constructorName', 'timeMillis', 'time', 'weather_class'],
      dtype='object')
```

▶ Code

```
season                0
round                 0
raceName              0
url                   0
circuitName           0
locality              0
country               0
date                  0
position              0
points                0
grid                  0
laps                  0
status                0
driverId              0
driverGivenName       0
driverFamilyName      0
constructorId         0
constructorName       0
timeMillis         1291
time               1291
weather_class         0
dtype: int64
```

The missing values in '`timeMillis'` and'`time'` columns are of those drivers who did not finish the race. Therefore, we will drop these columns and try to analyse the performance based on other metrics.

▶ Code

▶ Code

```
constructorName
Ferrari           235
McLaren           234
Williams          229
Red Bull          184
Renault           142
Sauber            139
Mercedes          134
Toro Rosso        127
Force India       100
Haas F1 Team       79
Toyota             76
Jordan             57
BAR                56
Minardi            54
Alfa Romeo         50
Jaguar             45
BMW Sauber         40
AlphaTauri         40
Lotus F1           38
Alpine F1 Team     30
Aston Martin       30
Honda              28
```

```
Arrows             26
HRT                24
Marussia           24
Super Aguri        24
Caterham           24
Racing Point       20
Prost              18
Benetton           18
Virgin             16
Manor Marussia     16
Lotus              16
Brawn              10
MF1                 9
Spyker              8
Name: count, dtype: int64
```

▶ Code

Some of the team names were changed in the process of rebranding or due to a change in ownership. For accurate analysis, we will replace the older versions of the constructors' names with the current ones.

- Ferrari
- McLaren
- Jaguar
- Williams
- Sauber → BMW Sauber → Sauber → Alfa Romeo → Sauber
- BAR → Honda → Brawn → Mercedes
- Benetton → Renault → Lotus F1 → Renault → Alpine
- Jordan → Midland → Spyker → Force India → Aston Martin
- Minardi → Toro Rosso → Scuderia AlphaTauri
- Haas
- Toyota
- Virgin Racing → Marussia F1 → Manor Marussia
- Lotus → Caterham
- Arrows
- Super Aguri
- HRT
- Prost

▶ Code

▶ Code

▶ Code

```
constructorName
Ferrari           235
McLaren           234
Red Bull          229
Williams          229
Sauber            229
Mercedes          228
Alpine F1 Team    228
```

```
Aston Martin      224
AlphaTauri        221
Haas F1 Team       79
Toyota             76
Marussia           56
Caterham           40
Arrows             26
Super Aguri        24
HRT                24
Prost              18
Name: count, dtype: int64
```

▶ Code

▶ Code

▶ Code

	season	round	raceName	url	circuit
0	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
1	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit

Classify the satus column in to broader categories. This column provides information on whether the driver has finished the race or not, if not, was it because of a mechanical failure, an accident, or was he lapped. The categories are:

- Finished
- Lap
- Accident
- Mechanical

Grouping the data helps understand the major reasons for the race results without getting overwhelmed by the granular details.

▶ Code

```
array(['Finished', '+1 Lap', '+2 Laps', 'Electrical', 'Hydraulics',
       'Overheating', 'Gearbox', 'Suspension', 'Accident', '+5
Laps',
       'Wheel', 'Engine', 'Spun off', 'Collision', '+3 Laps', '+4
Laps',
       '+10 Laps', 'Throttle', 'Clutch', 'Technical', 'Mechanical',
       'Driveshaft', 'Transmission', 'Steering', 'Puncture',
'Brakes',
       'Retired', 'Tyre', 'Fuel pressure', '+9 Laps', 'Water leak',
       'Disqualified', 'Did not qualify', '+42 Laps', 'Engine
misfire',
       'Power Unit', 'Oil pressure', 'Safety concerns', 'Fuel
system',
       '+6 Laps', 'Electronics', 'Collision damage', 'Wheel nut',
```

```
       'Battery', 'Oil leak', '+7 Laps', 'Stalled', 'Exhaust',
       'Vibrations', 'Broken wing', 'Fuel', 'Wheel rim', 'Power
loss',
       '107% Rule', '+8 Laps', 'ERS', 'Withdrew', 'Cooling system',
       'Water pump', 'Fuel leak', 'Front wing', 'Turbo', 'Damage',
       'Out of fuel', 'Radiator', 'Oil line', 'Fuel rig',
       'Launch control', 'Not classified', 'Pneumatics',
'Differential'],
      dtype=object)
```

▶ Code

▶ Code

```
status
Finished      1105
Lapped         693
Mechanical     412
Accident       190
Name: count, dtype: int64
```

▶ Code

Finish category - new categorical variable

In F1, race outcomes are categorized based on finishing positions:

- The top 3 finishers are celebrated on the podium and referred to as achieving a **Podium Finish**
- All drivers who finish in top 10 earn championship points, ranging from 25 points for the winner, 18 for second place, and decreasing to 1 point for the 10th position.
- Drivers finishing outside the top 10 do not earn any championship points.

▶ Code

▶ Code

	season	round	raceName	url	circuit
0	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
1	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
2	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit
3	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit

	season	round	raceName	url	circuit
4	2010	1	Bahrain Grand Prix	http://en.wikipedia.org/wiki/2010_Bahrain_Gran...	Bahrai Interna Circuit

▶ Code

Race Track Features

▶ Code

	Year	Grand Prix	Track Length (m)	Max Speed (km/h)	Full Throttle (%)	Number of Corners	Number of Straights
0	2020	Pre-Season Test 1	4312.438437	323	70.673953	1	4
1	2020	Pre-Season Test 2	4312.438437	323	70.673953	1	4
2	2020	Austrian Grand Prix	4312.438437	323	70.673953	1	4
3	2020	Styrian Grand Prix	4292.610384	300	46.556886	2	6
4	2020	Hungarian Grand Prix	4348.049386	318	58.114374	0	6

▶ Code

```
Year                  0
Grand Prix            0
Track Length (m)      0
Max Speed (km/h)      0
Full Throttle (%)     0
Number of Corners     0
Number of Straights   0
dtype: int64
```

Standardization

Standardization is a preprocessing step used to scale the features is a consistent scale for more accurate and stable modelling. Features with different scales can lead to one feature dominating others during model training, Standardization eliminates disparity.

For the Race Track Features, `StandardScaler` is used, this standardizes the data by “removing the mean and scaling to unit variance”[1].

Formula for Standardization:

$$Z = \frac{X - \mu}{\sigma}$$

where:

- X: the original value
- μ: mean
- σ: standard deviation of the feature

▶ Code

▶ Code

	Year	Grand Prix	Track Length (m)	Max Speed (km/h)	Full Throttle (%)	Number of Corners	Number of Straights
0	2020	Pre-Season Test 1	-1.000607	-0.115670	1.059667	-0.789651	-0.938394
1	2020	Pre-Season Test 2	-1.000607	-0.115670	1.059667	-0.789651	-0.938394
2	2020	Austrian Grand Prix	-1.000607	-0.115670	1.059667	-0.789651	-0.938394
3	2020	Styrian Grand Prix	-1.024865	-1.840980	-1.757479	-0.275003	-0.037811
4	2020	Hungarian Grand Prix	-0.957039	-0.490737	-0.407433	-1.304300	-0.037811

▶ Code

Pitstop data

▶ Code

	Year	Round	RaceName	DriverID	Lap	Stop	Time	Duration
0	2011	1	Australian Grand Prix	alguersuari	1	1	17:05:23	26.898
1	2011	1	Australian Grand Prix	michael_schumacher	1	1	17:05:52	25.021
2	2011	1	Australian Grand Prix	webber	11	1	17:20:48	23.426
3	2011	1	Australian Grand Prix	alonso	12	1	17:22:34	23.251
4	2011	1	Australian Grand Prix	massa	13	1	17:24:10	23.842

Pivoting involves reshaping data by rearranging rows into columns. It is used to transform long-format data (many rows for each entity) to wide-format (one row per entity with multiple columns).

Here, each `Stop` Number becomes a seperate set of columns, Lap1, Lap2, Duration1, Duration2 and so on.

▶ Code

	Year	Round	RaceName	DriverID	Lap1	Lap2	Lap3	Lap4	Lap5	Lap6	...	Time
0	2011	1	Australian Grand Prix	alguersuari	1	17	35	0	0	0	...	0
1	2011	1	Australian Grand Prix	alonso	12	27	42	0	0	0	...	0

2 rows × 32 columns

▶ Code

```
(5118, 32)
```

MinMaxx Scaling: Transforms the fature range by scaling its min and max values.

Formula:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

▶ Code

```
   Year  Round               RaceName     DriverID      Lap1
Lap2  \
0  2011      1  Australian Grand Prix  alguersuari  0.000000
0.229730
1  2011      1  Australian Grand Prix       alonso  0.174603
0.364865
2  2011      1  Australian Grand Prix     ambrosio  0.206349
0.513514
3  2011      1  Australian Grand Prix  barrichello  0.190476
0.310811
4  2011      1  Australian Grand Prix        buemi  0.222222
0.391892

       Lap3      Lap4  Lap5  Lap6  ...  Time5  Time6  Time7
Duration1  \
0  0.479452  0.000000   0.0   0.0  ...    0.0    0.0    0.0
0.453661
1  0.575342  0.000000   0.0   0.0  ...    0.0    0.0    0.0
0.392151
2  0.000000  0.000000   0.0   0.0  ...    0.0    0.0    0.0
0.426017
3  0.383562  0.512821   0.0   0.0  ...    0.0    0.0    0.0
0.398762
4  0.000000  0.000000   0.0   0.0  ...    0.0    0.0    0.0
0.427417

   Duration2  Duration3  Duration4  Duration5  Duration6  Duration7
0   0.428042   0.457423   0.000000        0.0        0.0        0.0
1   0.432766   0.419802   0.000000        0.0        0.0        0.0
2   0.462739   0.000000   0.000000        0.0        0.0        0.0
3   0.662386   0.293259   0.469108        0.0        0.0        0.0
4   0.404192   0.000000   0.000000        0.0        0.0        0.0
```

```
[5 rows x 32 columns]
```

▶ Code

All the clean datasets can be found here

Footnotes

1. StandardScaler ↩