Data Transformations for Feature Analysis

Date and Time in Python

Python provides a module named `datetime` to deal with dates and times.
It allows you to set `date` , `time` or both `date` and `time` using the `date()` , `time()` and `datetime()` functions respectively, after importing the `datetime` module .

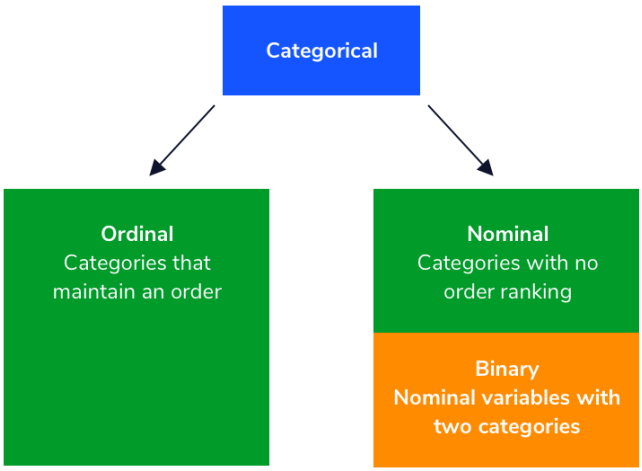
```
import datetime
feb_16_2019 = datetime.date(year=2019,
month=2, day=16)
feb_16_2019 = datetime.date(2019, 2, 16)
print(feb_16_2019) #2019-02-16

time_13_48min_5sec = datetime.time(hour=13,
minute=48, second=5)
time_13_48min_5sec = datetime.time(13, 48, 5)
print(time_13_48min_5sec) #13:48:05

timestamp= datetime.datetime(year=2019,
month=2, day=16, hour=13, minute=48,
second=5)
timestamp = datetime.datetime(2019, 2, 16,
13, 48, 5)
print (timestamp) #2019-01-02 13:48:05
```

Categorical Variables

Categorical variables consist of data that can be grouped into distinct categories, and are ordinal or nominal. Ordinal categorical variables which are groups that contain an inherent ranking, such as ratings of plays or responses to a survey question with a point scale e.g., on a scale from 1-7, how happy are you right now? Nominal categorical variables are made of categories without an inherent order, examples of nominal variables are species of ants, or people's hair color.



One-Hot Encoding with Python

When working with nominal categorical variables in Python, it can be useful to use One-Hot Encoding, which is a technique that will effectively create binary variables for each of the nominal categories. This encodes the variable without creating an order among the categories. To one-hot encode a variable in a pandas dataframe, we can use the `.get_dummies()`.

```
df = pd.get_dummies(data = df, columns=
['column1', 'column2')
```

Categorical Data Frequencies

One way to summarize a categorical variable is to compute the frequencies of the categories. For further summarization, the frequency of the modal category (most frequent category) is often reported. For example, when analyzing a dataset with an education level variable (highschool, associates, bachelors, masters, etc.), we could calculate the frequency of each category and report the most common category. For a pandas dataframe, we can use the `.value_counts()` method on a column of data to calculate the frequencies of the categories.

```
# calculate counts of values for a column in
a dataframe:
df['column_name'].value_counts()
```

Categorical Data Defined

Categorical Data refers to data represented by words rather than numbers. Examples of categorical data are tree species and survey responses (Agree, Neutral, Disagree).

Ordinal and Nominal Categorical Data

Categorical variables can be either ordinal (ordered) or nominal (unordered).
Examples of ordinal variables include places (1st, 2nd, 3rd) and survey responses (on a scale of 1 to 5, how much do you agree with a statement).
Examples of nominal variables include tree species, student names, and account names.