***Writing a Data Description Report***

***To proceed effectively with your data mining project, consider the value of producing an accurate data description report using the following metrics:***

***Data Quantity***

***• What is the format of the data?***

* ***The data is given in csv file format. in order to access it var file can be used.***

***• Identify the method used to capture the data--for example, ODBC.***

* ***As of now the method is not clearly known.***

***• How large is the database (in numbers of rows and columns)?***

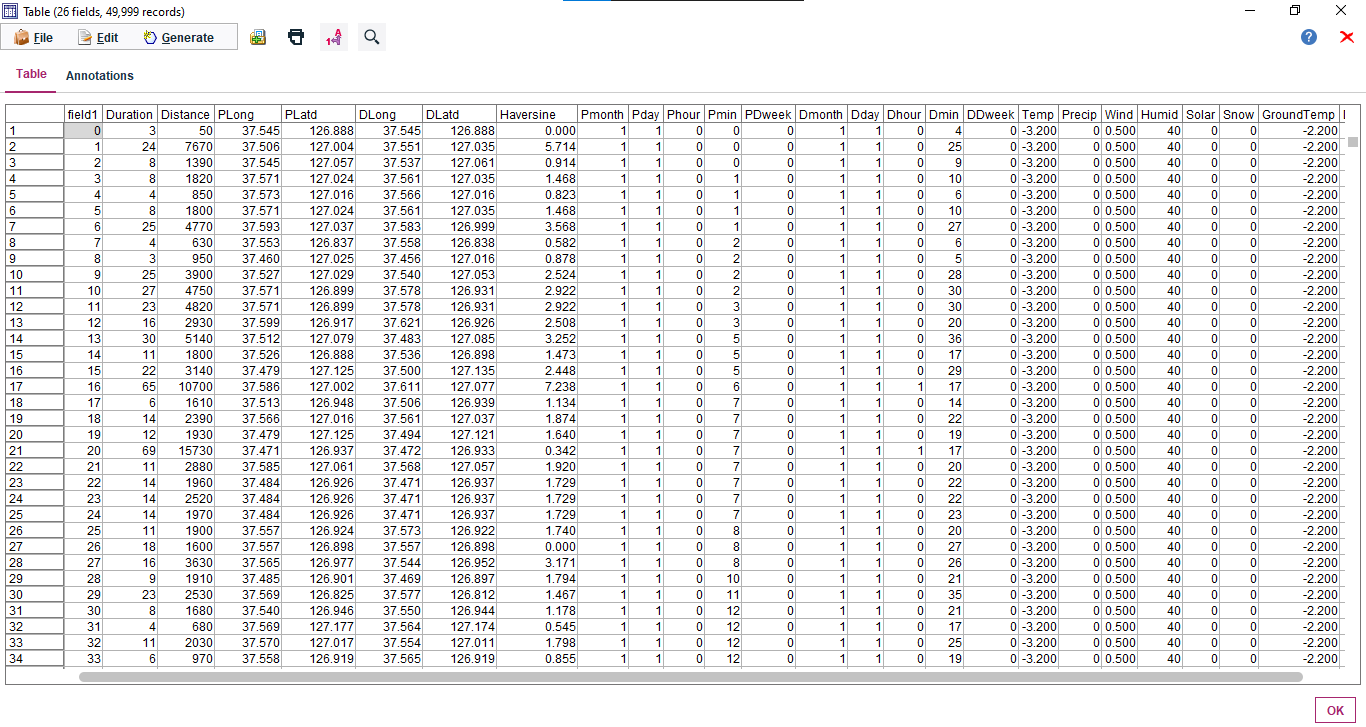
***For\_modeling.csv:***

***File size: 1.2 GB***

***Rows: 9,601,139***

***Columns: 26***

**Brief about the dataset:**

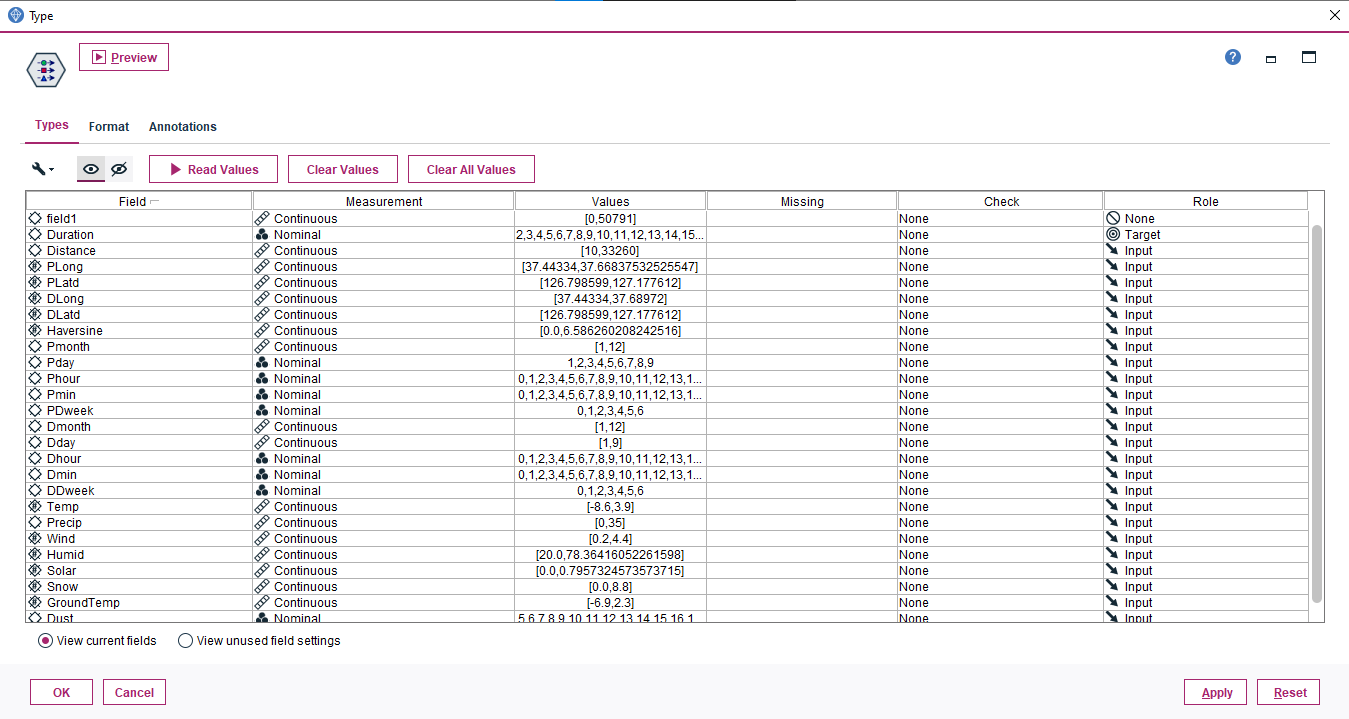
******

**Exploration Model:**

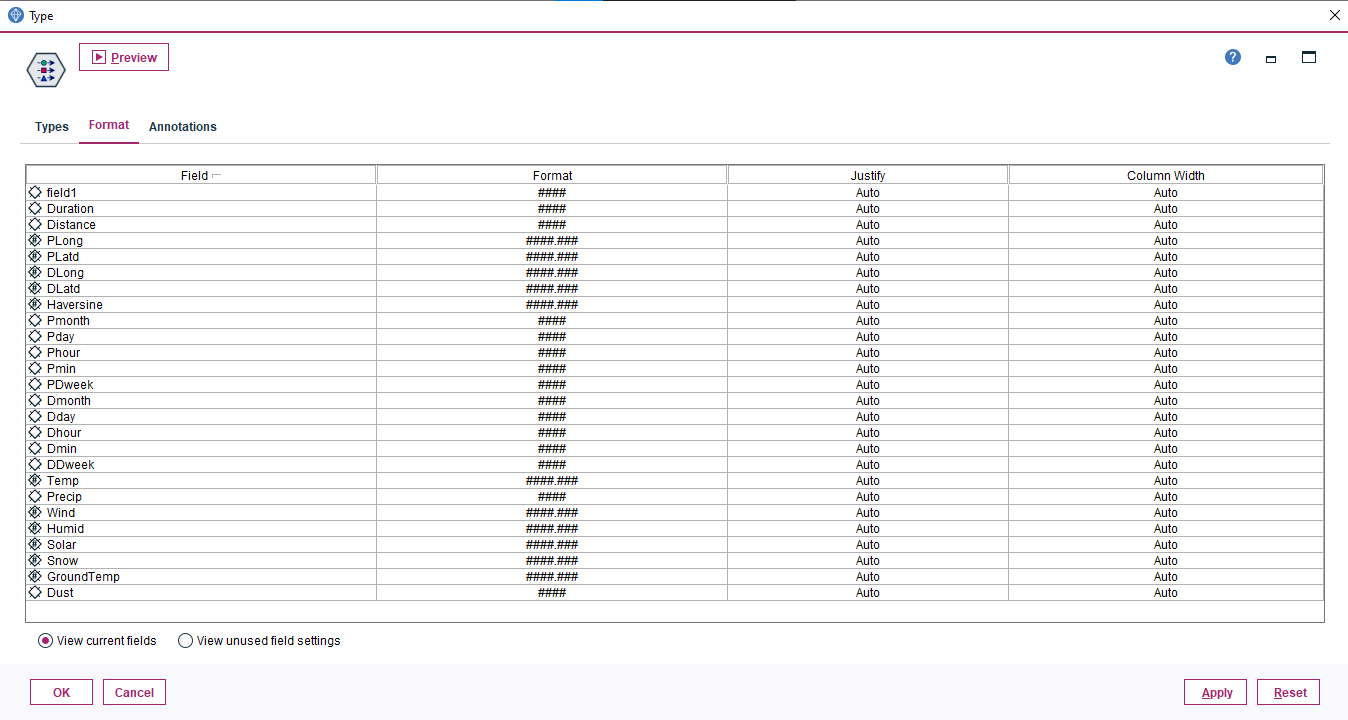
The model used to describe the data and gain insights on its nature. Please note that the target as discussed before has been set to the Target and Failure feild

Information about the input variables

The file T**est.csv** provides the details of its customers including:

******

**Format of each field:**

******

Looking into each field and its attributes

1. **Unnamed – which is ID.**

Measurement Level = It is type less, data that doesn't conform to any of the above types can be described as a set with many members

Its data type is string as it has alpha-numeric values

Role = It is automatically set to none; since measurement level is type less this will not be a contributing factor to output but just an identifier

Format = No format specified

1. **Duration – which is Our Target Column**

*Measurement Level* = Nominal indicating that it has distinct values set within the range 1 to 119

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values

1. **Distance**

*Measurement Level* = Continuous indicating that it has continuous values within the range 1 to 33290

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values

1. **PLong**

*Measurement Level* = Continuous indicating that it has continuous values within the range 37.437271 to 37.68972

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **PLatd**

*Measurement Level* = Continuous indicating that it has continuous values within the range 126.798599 to 127.180267

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Dlong**

*Measurement Level* = Continuous indicating that it has continuous values within the range 37.437271to 37.68972

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Dlatd**

*Measurement Level* = Continuous indicating that it has continuous values within the range 126.798599 to 127.180267

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Haversine**

*Measurement Level* = Continuous indicating that it has continuous values within the range 0.0 to 6.586260208242516

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Pmonth**

*Measurement Level* = Continuous indicating that it has continuous values within the range 1 to 12

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values

1. **PDay**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 1-9

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **PHour**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-23

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **PMin**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-59

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **PDweek**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-6

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **Dmonth**

*Measurement Level* = Continuous indicating that it has continuous values within the range 1 to 12

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values

1. **DDay**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 1-9

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **DHour**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-23

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **DMin**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-59

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **DDweek**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 0-6

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

1. **Temp**

*Measurement Level* = Continuous indicating that it has continuous values within the range -8.6 to 3.9

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Precip**

*Measurement Level* = Continuous indicating that it has continuous values within the range 0 to 35

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Wind**

*Measurement Level* = Continuous indicating that it has continuous values within the range 0.2 to 4.4

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Humid**

*Measurement Level* = Continuous indicating that it has continuous values within the range 20.0 to 78.36416052261598

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Solar**

*Measurement Level* = Continuous indicating that it has continuous values within the range 0.0 to 0.7957324573573715

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Snow**

*Measurement Level* = Continuous indicating that it has continuous values within the range 0.0 to 8.8

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **GroundTemp**

*Measurement Level* = Continuous indicating that it has continuous values within the range -6.9 to 2.3

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####.###, indicates that this column has maximum 4 digits and 3 decimal places in its integer values

1. **Dust**

*Measurement Level* = It is Nominal, used to describe data with multiple distinct values, each treated as a member of a set within range of 5-80

Its *data type* is Integer indicating that it only has Integer values.

*Role* = Input, it means that this column/field is a contributing factor in determining the output

Format = ####, indicates that this column has maximum 4 digits in its integer values.

**Brief glossary of the datatypes for reference:**

The following measurement levels are available:

* **Default.** Data whose storage type and values are unknown (for example, because they haven't yet been read) are displayed as **Default**.
* **Continuous.** Used to describe numeric values, such as a range of 0–100 or 0.75–1.25. A continuous value can be an integer, real number, or date/time.
* **Categorical.** Used for string values when an exact number of distinct values is unknown. This is an *uninstantiated* data type, meaning that all possible information about the storage and usage of the data is not yet known. Once data is read, the measurement level will be **Flag**, **Nominal**, or **Typeless**, depending on the maximum number of members for nominal fields specified.
* **Flag.** Used for data with two distinct values that indicate the presence or absence of a trait, such as true and false, Yes and No, or 0 and 1. The values used may vary, but one must always be designated as the "true" value, and the other as the "false" value. Data may be represented as text, integer, real number, date, time, or timestamp.
* **Nominal.** Used to describe data with multiple distinct values, each treated as a member of a set, such as small/medium/large. Nominal data can have any storage—numeric, string, or date/time. Note that setting the measurement level to Nominal doesn't automatically change the values to string storage.
* **Ordinal.** Used to describe data with multiple distinct values that have an inherent order. For example, salary categories or satisfaction rankings can be typed as ordinal data. The order is defined by the natural sort order of the data elements. For example, 1, 3, 5 is the default sort order for a set of integers, while HIGH, LOW, NORMAL (ascending alphabetically) is the order for a set of strings. The ordinal measurement level enables you to define a set of categorical data as ordinal data for the purposes of visualization, model building, and export to other applications (such as IBM SPSS Statistics) that recognize ordinal data as a distinct type. You can use an ordinal field anywhere that a nominal field can be used. Additionally, fields of any storage type (real, integer, string, date, time, and so on) can be defined as ordinal.
* **Typeless.** Used for data that doesn't conform to any of the above types, for fields with a single value, or for nominal data where the set has more members than the defined maximum. It's also useful for cases in which the measurement level would otherwise be a set with many members (such as an account number). When you select Typeless for a field, the role is automatically set to None, with Record ID as the only alternative. The default maximum size for sets is 250 unique values.