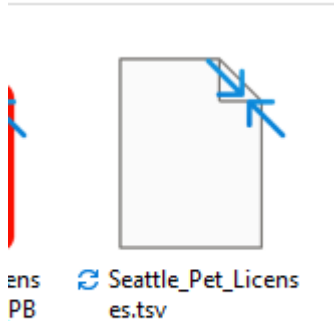# Alteryx - Perform Data profile and prepration on Seattle Pet License
Submitted by Yamini Manral, NUID 002703881

Upload all execution screen shots which should contain -

1. Download Seattle pet license data from Seattle Pet Licenses | City of Seattle Open Data portal (TSV format)
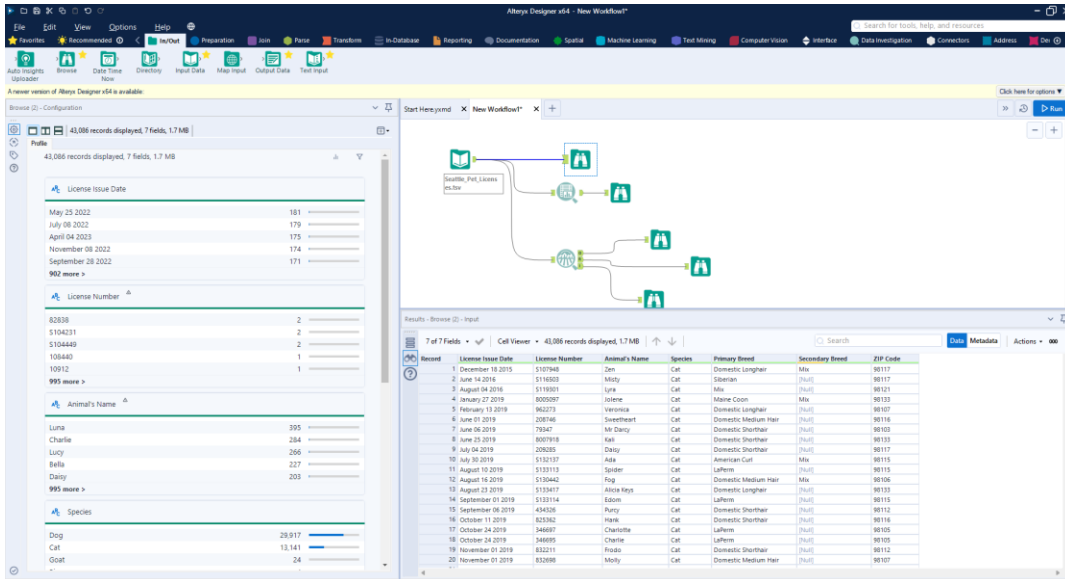

Loaded TSV file into Alteryx:

2. Profile metrics and your analysis description the analysis

After loading the data, Alteryx correctly displays the number of records and columns for the input TSV file.

Some useful insights that we can immediately derive from the initial profiling:

- Most popular pet name is LUNA
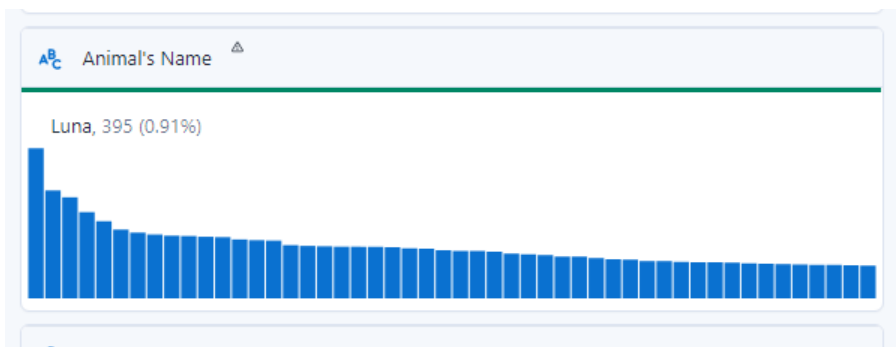
Frequent Values	
Luna	395
Charlie	284
Lucy	266
Bella	227
Daisy	203
Max	181

Summary	
Count	43086
Valid	43054
Unique	12780
Blank	32
Longest	Rose Lee the Huntress Wilderness Spirit Wand
Shortest	Z

- Of all the pets owned, DOGS is the most common pet species

Species

Dog, 29,917 (69.43%)

---

1 record displayed, 1 field, 34 KB

Report   Profile

1 of 1 Fields ▾ ✓ | ◄◄ ◄   Records 1 to 1   ► ►►|

Species

Frequent Values

Dog	29917
Cat	13141
Goat	24
Pig	4

Summary

Count	43086
Valid	43086
Unique	4
Blank	0
Longest	Goat
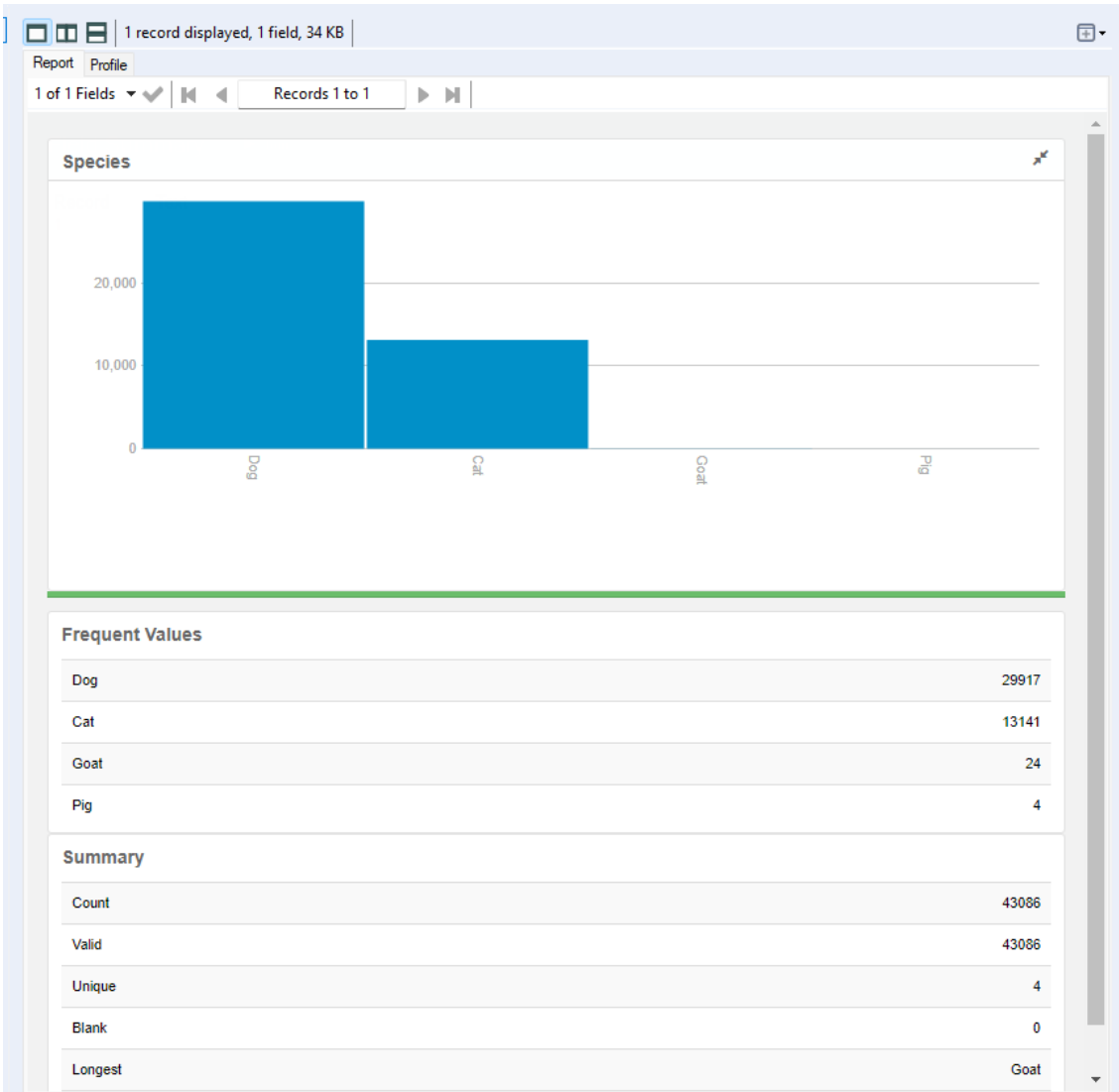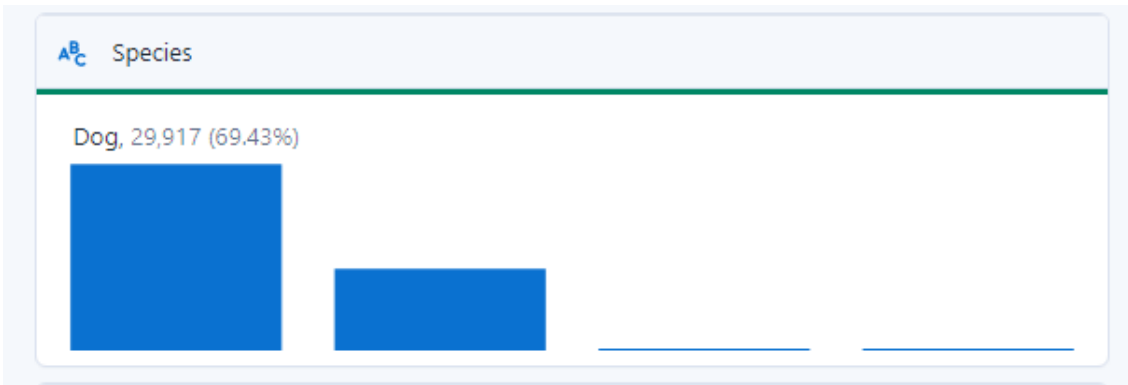
- Most popular Primary breed is DOMESTIC SHORTHAIR (Cats)

Primary Breed

Domestic Shorthair, 7,307 (16.95%)

Primary Breed

Frequent Values

Domestic Shorthair	7307
Retriever, Labrador	3309
Domestic Medium Hair	1589
Retriever, Golden	1524
Chihuahua, Short Coat	1502
American Shorthair	1158

Summary

Count	43086
Valid	43086
Unique	341
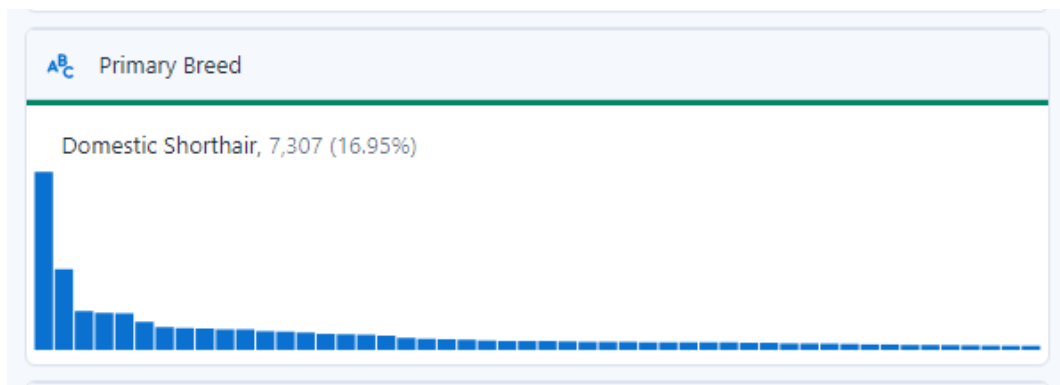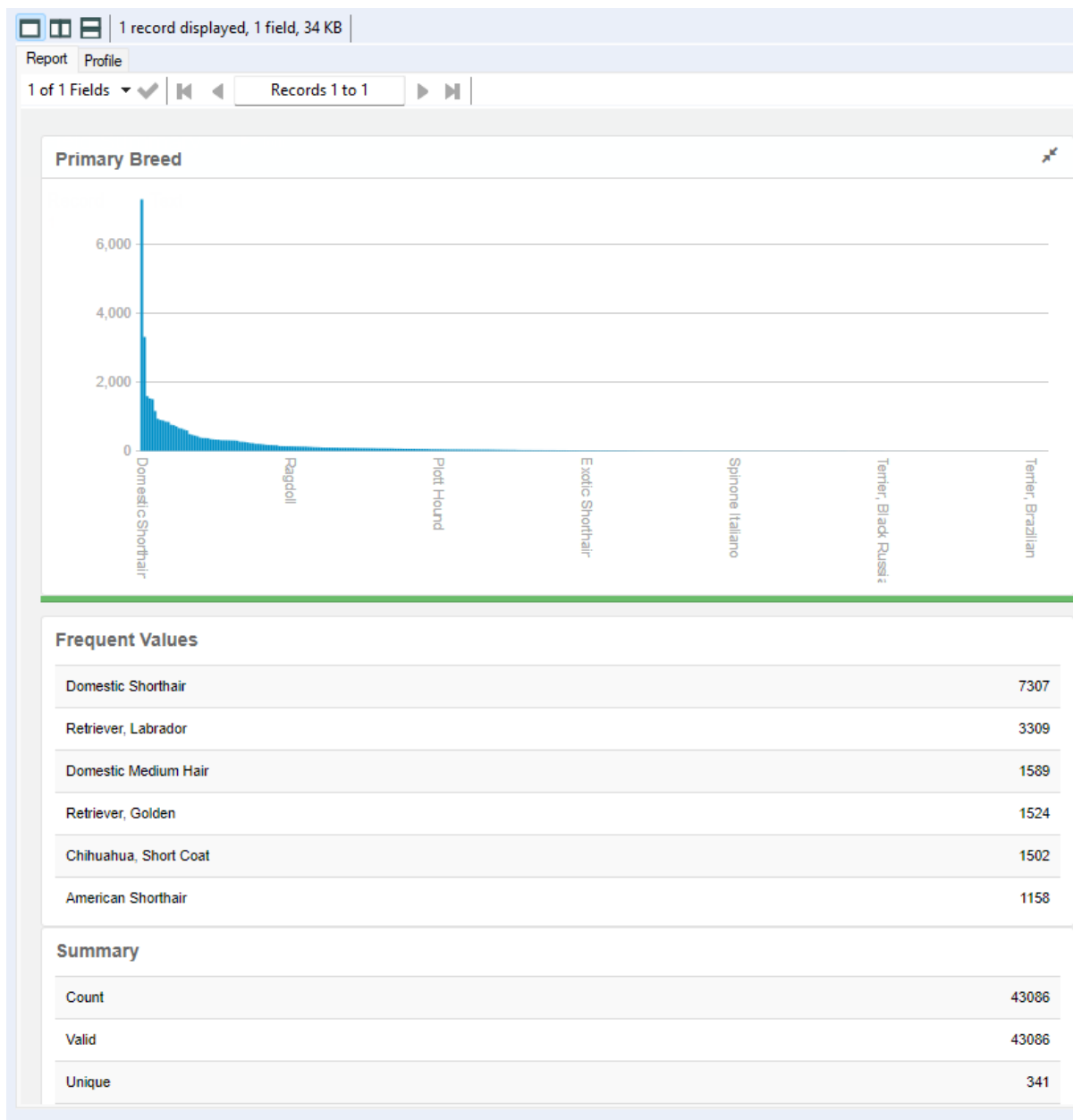
- Most common secondary breed is MIX

Secondary Breed

Mix, 17,458 (40.51%)
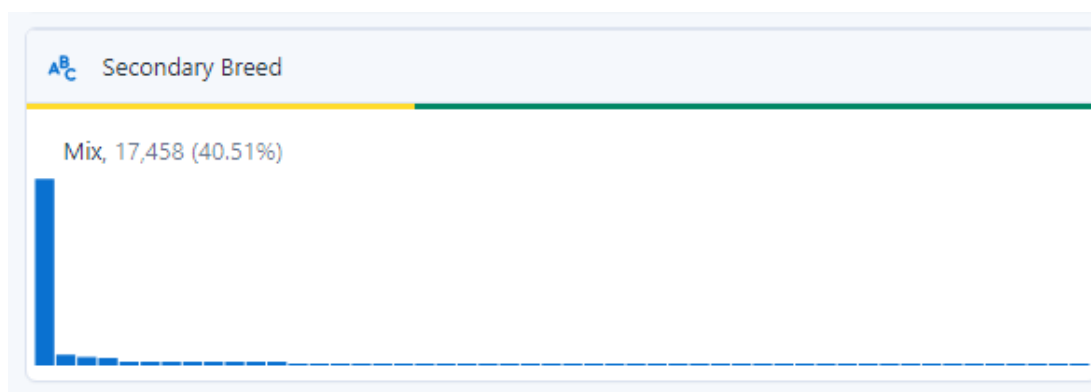
- Most pets are owned in the area with ZIP 98115

ABC ZIP Code

98115, 4,038 (9.37%)

ZIP Code

4,000

2,000

0

98115   98027   98165   31082   96740   98292   98328

Frequent Values

98115	4038
98103	3689
98117	3100
98125	2416
98122	2151
98118	2084
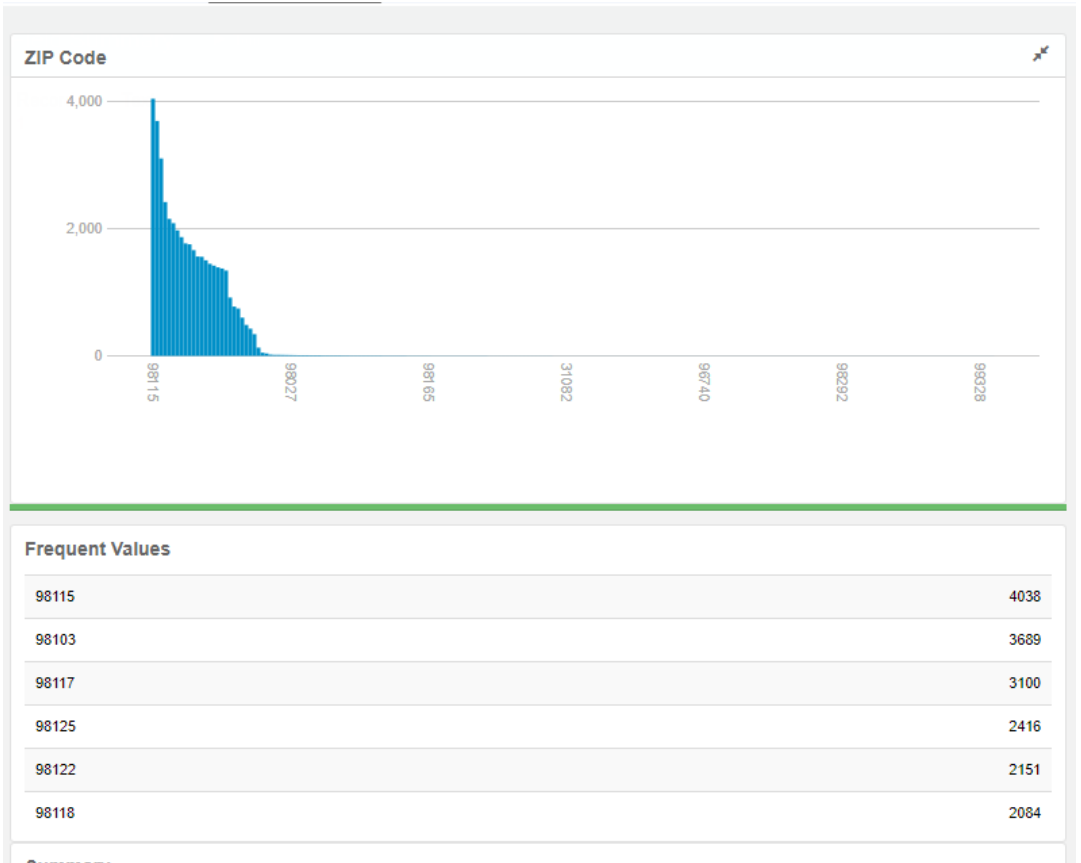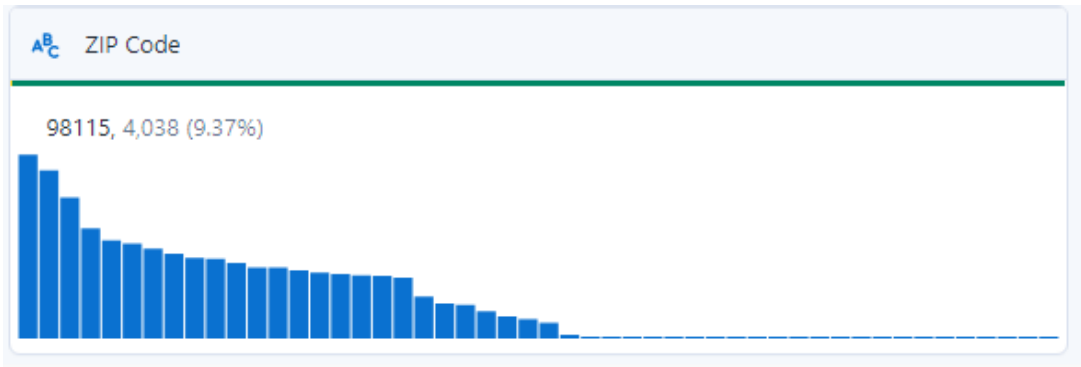
Basic Profiling correctly displays features of data for all the 7 fields:

Record	FieldName	Name	Value
1	License Issue Date	Name	License Issue Date
2	License Issue Date	Data Type	V_WString
3	License Issue Date	Size	254
4	License Issue Date	Source	Formula: [_CurrentField_]
5	License Issue Date	Description	[Null]
6	License Issue Date	OKs	43086
7	License Issue Date	Nulls	0
8	License Issue Date	Non-Nulls	43086
9	License Issue Date	Blanks	0
10	License Issue Date	Values with Leading Whitespace	0
11	License Issue Date	Values with Trailing Whitespace	0
12	License Issue Date	Values with Both Whitespace	0
13	License Issue Date	Average Length	14.2
14	License Issue Date	Longest Length	17
15	License Issue Date	Longest Value	September 01 2019
16	License Issue Date	Shortest (Non Blank) Length	11
17	License Issue Date	Shortest Value	May 19 2020
18	License Issue Date	Minimum	April 01 2021
19	License Issue Date	Maximum	September 30 2022
20	License Issue Date	Uniques	907
21	License Issue Date	Unique Values	[Null]
22	License Number	Name	License Number
23	License Number	Data Type	V_WString
24	License Number	Size	254
25	License Number	Source	Formula: [_CurrentField_]
26	License Number	Description	[Null]
27	License Number	OKs	43086
28	License Number	Nulls	0
29	License Number	Non-Nulls	43086
30	License Number	Blanks	0
31	License Number	Values with Leading Whitespace	0
32	License Number	Values with Trailing Whitespace	0
33	License Number	Values with Both Whitespace	0
34	License Number	Average Length	6.8
35	License Number	Longest Length	7
36	License Number	Longest Value	S107948
37	License Number	Shortest (Non Blank) Length	4
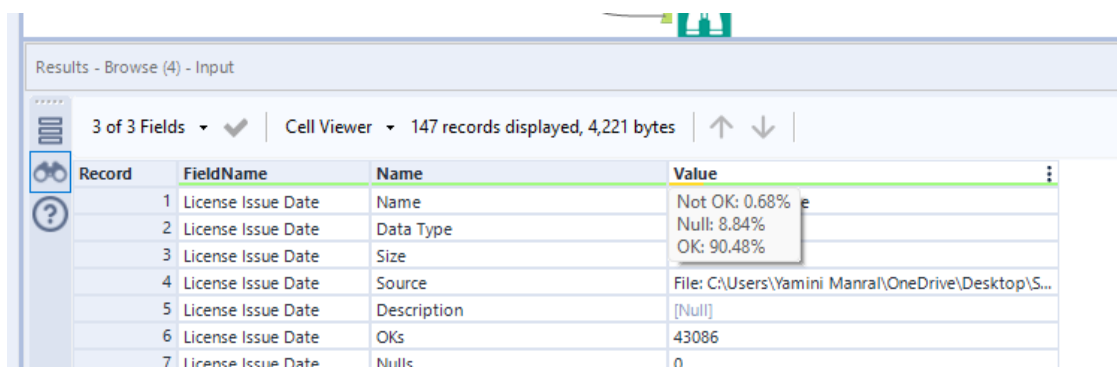
Record	FieldName	Name	Value
56	Animal's Name	Longest Length	50
57	Animal's Name	Longest Value	Nuit Ahathoor Hecate Sapph...
58	Animal's Name	Shortest (Non Blank) Length	1
59	Animal's Name	Shortest Value	Z
60	Animal's Name	Minimum	"Zippy"
61	Animal's Name	Maximum	"Dr." Noah Pickles
62	Animal's Name	Uniques	> 10000
63	Animal's Name	Unique Values	[Null]
64	Species	Name	Species
65	Species	Data Type	V_WString
66	Species	Size	254
67	Species	Source	Formula: [_CurrentField_]
68	Species	Description	[Null]
69	Species	OKs	43086
70	Species	Nulls	0
71	Species	Non-Nulls	43086
72	Species	Blanks	0
73	Species	Values with Leading Whitespace	0
74	Species	Values with Trailing Whitespace	0
75	Species	Values with Both Whitespace	0
76	Species	Average Length	3.0
77	Species	Longest Length	4
78	Species	Longest Value	Goat
79	Species	Shortest (Non Blank) Length	3
80	Species	Shortest Value	Cat
81	Species	Minimum	Cat
82	Species	Maximum	Pig
83	Species	Uniques	4
84	Species	Unique Values	Cat
85	Primary Breed	Name	Primary Breed
86	Primary Breed	Data Type	V_WString
87	Primary Breed	Size	254
88	Primary Breed	Source	Formula: [_CurrentField_]
89	Primary Breed	Description	[Null]
90	Primary Breed	OKs	43086
91	Primary Breed	Nulls	0
92	Primary Breed	Non-Nulls	43086

Record	FieldName	Name	Value
107	Secondary Breed	Data Type	V_WString
108	Secondary Breed	Size	254
109	Secondary Breed	Source	Formula: [_CurrentField_]
110	Secondary Breed	Description	[Null]
111	Secondary Breed	OKs	27425
112	Secondary Breed	Nulls	15661
113	Secondary Breed	Non-Nulls	27425
114	Secondary Breed	Blanks	0
115	Secondary Breed	Values with Leading Whitespace	0
116	Secondary Breed	Values with Trailing Whitespace	0
117	Secondary Breed	Values with Both Whitespace	0
118	Secondary Breed	Average Length	8.0
119	Secondary Breed	Longest Length	46
120	Secondary Breed	Longest Value	Mixed Breed, Medium (up to…
121	Secondary Breed	Shortest (Non Blank) Length	3
122	Secondary Breed	Shortest Value	Mix
123	Secondary Breed	Minimum	Abyssinian
124	Secondary Breed	Maximum	York Chocolate
125	Secondary Breed	Uniques	262
126	Secondary Breed	Unique Values	[Null]
127	ZIP Code	Name	ZIP Code
128	ZIP Code	Data Type	V_WString
129	ZIP Code	Size	254
130	ZIP Code	Source	Formula: [_CurrentField_]
131	ZIP Code	Description	[Null]
132	ZIP Code	OKs	42957
133	ZIP Code	Nulls	129
134	ZIP Code	Non-Nulls	42957
135	ZIP Code	Blanks	0
136	ZIP Code	Values with Leading Whitespace	0
137	ZIP Code	Values with Trailing Whitespace	0
138	ZIP Code	Values with Both Whitespace	0
139	ZIP Code	Average Length	5.0
140	ZIP Code	Longest Length	5
141	ZIP Code	Longest Value	98117
142	ZIP Code	Shortest (Non Blank) Length	3
143	ZIP Code	Shortest Value	981

Which further shows our data has been loaded correctly.

3. Document what can be done to fix/filter out the bad data (no need to perform any code on this. Analysis is enough for this assignment submission)

Results - Browse (4) - Input

3 of 3 Fields ▾ ✓ | Cell Viewer ▾ 147 records displayed, 4,221 bytes | ↑ ↓ |

Record	FieldName	Name	Value
1	License Issue Date	Name	Not OK: 0.68% e
2	License Issue Date	Data Type	Null: 8.84%
3	License Issue Date	Size	OK: 90.48%
4	License Issue Date	Source	File: C:\Users\Yamini Manral\OneDrive\Desktop\S...
5	License Issue Date	Description	[Null]
6	License Issue Date	OKs	43086
7	License Issue Date	Nulls	0

There are a lot of NULL values in secondary breed column, if we remove it and set it as blank/undefined (a string), it takes its own metrics in data profiling, but the data is cleaner now; but if we don't, then it shows up as a warning as NULL values are not encouraged and does not interrupt with the data profiling results.

4. execution runtime (how long does it take to complete the workflow execution)

Results - Workflow - Messages

All | ⓘ 0 Errors | ⓘ 3 Conv Errors | ⚠ 0 Warnings | ☐ 2 Messages | 🖬 6 Files | Last Run | Configuration

	Designer x64	Started running C:\Users\Yamini Manral\OneDrive\Desktop\Summer Semester\DAMG7370 DADABI\Assignments - DADABI\A3\DataProfiling_Seattl
	Designer x64	The Designer x64 reported: Allocating requested memory would be more than available physical memory. Reverting to 1824.3 MB of memory.
	Input Data (10)	Autodetected Code Page: UTF-8
	Input Data (10)	43086 records were read from "C:\Users\Yamini Manral\OneDrive\Desktop\Summer Semester\DAMG7370 DADABI\Assignments - DADABI\A3\Seattle_Pet_Licenses.tsv"
	Browse (4)	147 records
ⓘ	Field Summary	Tool #275: Tool #32: Value: "Agnes "Aggie" Snowball Pineapple Clark" was truncated
ⓘ	Field Summary	Tool #275: Tool #32: Value: "Cascade Mountain's Out of a Dream (BRIA)" was truncated
ⓘ	Field Summary	Tool #275: Tool #32: Value: "Houlene "Hula" Elizabeth Honeybee Jacobs" was truncated
	Browse (7)	1 records
	Browse (8)	1 records
	Browse (6)	7 records
	Browse (2)	43,086 records
ⓘ	Designer x64	Finished running DataProfiling_SeattlePetLicensing_manral.y.yxmd in 3.1 seconds with 3 field conversion errors

Browse (2)   43,086 records

ⓘ Designer x64   Finished running DataProfiling_SeattlePetLicensing_manral.y.yxmd in 3.1 seconds with 3 field conversion errors

All | ⓘ 0 Errors | ⓘ 3 Conv Errors | ⚠ 0 Warnings | ☐ 0 Messages | 🖬 0 Files

ⓘ Field Summary (5) | Tool #275: Tool #32: Value: "Agnes "Aggie" Snowball Pineapple Clark" was truncated

ⓘ Field Summary (5) | Tool #275: Tool #32: Value: "Cascade Mountain's Out of a Dream (BRIA)" was truncated

ⓘ Field Summary (5) | Tool #275: Tool #32: Value: "Houlene "Hula" Elizabeth Honeybee Jacobs" was truncated

5. number of rows inserted vs number of rows on TSV

What's in this Dataset?

Rows	Columns	Each row is a
43.1K	7	**Pet License**

Results - Browse (2) - Input

7 of 7 Fields ▾ ✓ | Cell Viewer ▾ 43,086 records displayed, 1.7 MB ↑ ↓ | 🔍 Search

Record	License Issue Date	License Number	Animal Name	Species	Primary Breed	Secondary Breed	ZIP Code
1	December 18 2015	S107948	Zen	Cat	Domestic Longhair	Mix	98117
2	June 14 2016	S116503	Misty	Cat	Siberian	[Null]	98117
3	August 04 2016	S119301	Lyra	Cat	Mix	[Null]	98121
4	January 27 2019	8005097	Jolene	Cat	Maine Coon	Mix	98133
5	February 13 2019	962273	Veronica	Cat	Domestic Longhair	[Null]	98107

```
April 07 2023    8017262    Theo  Dog   Shih Tzu     Terrier, Yorkshire    98118
April 07 2023    8045573    Bronny       Dog    Retriever   Mix   98105
April 07 2023    8045574    Red    Dog   Retriever    Mix    98105
April 07 2023    8045577    Linus Dog   German Shepherd    Mix    98146
```

Ln 43087, Col 58

6. Create 2 new columns
   o DI_Created_Date and store current system date and time
   o Surrogate key (Record_ID)

7. Screen shots of workflow execution

8. Screen shot from SSMS showing the table creation , its datatypes mapped appropriately and output of count rows on SQL Server in the newly created database.

Table imported in SSMS within Seattle_Pet_Licensing database, within user damg7370, with table name SPL_2023

	Record_ID	License Issue Date	License Number	Animal's Name	Species	Primary Breed	Secondary Breed	ZIP Code	DI_Created_Date
1	1	December 18 2015	S107948	Zen	Cat	Domestic Longhair	Mix	98117	2023-05-28 16:22:01
2	2	June 14 2016	S116503	Misty	Cat	Siberian	NULL	98117	2023-05-28 16:22:01
3	3	August 04 2016	S119301	Lyra	Cat	Mix	NULL	98121	2023-05-28 16:22:01
4	4	January 27 2019	8005097	Jolene	Cat	Maine Coon	Mix	98133	2023-05-28 16:22:01
5	5	February 13 2019	962273	Veronica	Cat	Domestic Longhair	NULL	98107	2023-05-28 16:22:01
6	6	June 01 2019	208746	Sweetheart	Cat	Domestic Medium Hair	NULL	98116	2023-05-28 16:22:01
7	7	June 06 2019	79347	Mr Darcy	Cat	Domestic Shorthair	NULL	98103	2023-05-28 16:22:01
8	8	June 25 2019	8007918	Kali	Cat	Domestic Shorthair	NULL	98133	2023-05-28 16:22:01
9	9	July 04 2019	209285	Daisy	Cat	Domestic Shorthair	NULL	98117	2023-05-28 16:22:01
10	10	July 30 2019	S132137	Ada	Cat	American Curl	Mix	98115	2023-05-28 16:22:01
11	11	August 10 2019	S133113	Spider	Cat	LaPerm	NULL	98115	2023-05-28 16:22:01
12	12	August 16 2019	S130442	Fog	Cat	Domestic Medium Hair	Mix	98106	2023-05-28 16:22:01
13	13	August 23 2019	S133417	Alicia Keys	Cat	Domestic Longhair	NULL	98133	2023-05-28 16:22:01
14	14	September 01 2019	S133114	Edom	Cat	LaPerm	NULL	98115	2023-05-28 16:22:01
15	15	September 06 2019	434326	Purcy	Cat	Domestic Shorthair	NULL	98112	2023-05-28 16:22:01
16	16	October 11 2019	825362	Hank	Cat	Domestic Shorthair	NULL	98116	2023-05-28 16:22:01
17	17	October 24 2019	346697	Charlotte	Cat	LaPerm	NULL	98105	2023-05-28 16:22:01
18	18	October 24 2019	346695	Charlie	Cat	LaPerm	NULL	98105	2023-05-28 16:22:01
19	19	November 01 2019	832211	Frodo	Cat	Domestic Shorthair	NULL	98112	2023-05-28 16:22:01
20	20	November 01 2019	832698	Molly	Cat	Domestic Medium Hair	NULL	98107	2023-05-28 16:22:01

Query executed successfully.

```
/****** Script for SelectTopNRows command from SSMS ******/
SELECT TOP (1000) [Record_ID]
      ,[License Issue Date]
      ,[License Number]
      ,[Animal's Name]
      ,[Species]
      ,[Primary Breed]
      ,[Secondary Breed]
      ,[ZIP Code]
      ,[DI_Created_Date ]
  FROM [Seattle_Pet_License].[dbo].[SPL_2023]
```

Total rows:

```
SELECT COUNT(*) AS total_rows
FROM dbo.SPL_2023;
```

150 %

	total_rows
1	43086

Data types:

```
SELECT COLUMN_NAME, DATA_TYPE
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_NAME = 'SPL_2023';
```

150 %

	COLUMN_NAME	DATA_TYPE
1	Record_ID	int
2	License Issue Date	char
3	License Number	char
4	Animal's Name	nvarchar
5	Species	char
6	Primary Breed	varchar
7	Secondary Breed	varchar
8	ZIP Code	int
9	DI_Created_Date	nvarchar