FIT5196 Reflective Report in Assessment 2

Student Name: Anushka Jemima

Student ID: 33617457

Student Name: Shruthi Shashidhara Shastry

Student ID: 33684669

Date: 18-10-2024

1. Dirty Data

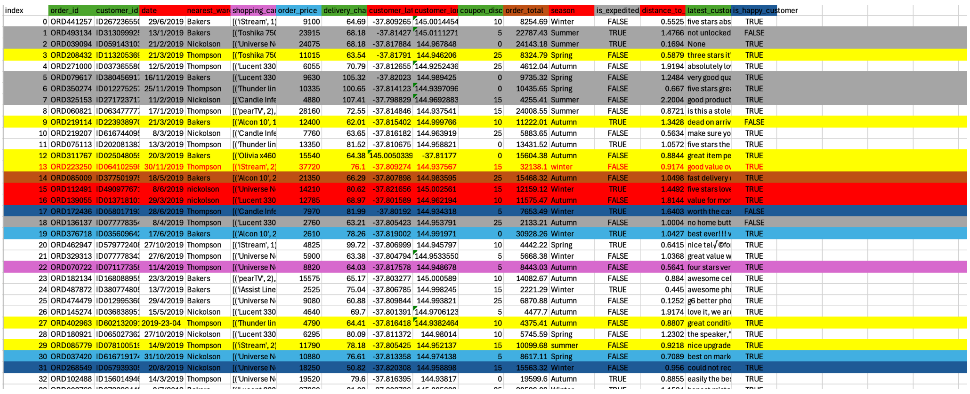
Feedback provided :

a. Order of finding and fixing anomalies is important.
b. IS_EXPIDITED column is related to delivery charge.
c. There can be more than one anomaly present in a column.

Improvements after Feedback :

After the feedback session we began examining the order in which we had discovered and fixed anomalies, we noticed for season and nearest warehouses the only anomaly was not the capitalization issue but there was mismatch between the season and months , and the nearest warehouse based on minimum distance.

We also ensured that we are first fixing the customer latitude and longitude column anomaly before the nearest warehouse as this caused 2 anomalies in the same row if we did the nearest warehouse first and then the customer latitude and longitude , to track this we added an index during exploration and maintained an excel file to ensure we have only one anomaly per row.



index	order_id	customer_id	date	nearest_ware	shopping_cai	order_price	delivery_cha	customer_lat	customer_lo	coupon_disc	order_total	season	is_expedited	distance_to_	latest_custo	is_happy_customer
0	ORD441257	ID26723655(	29/6/2019	Bakers	[('iStream', 1)	9100	64.69	-37.809265	145.0014454	10	8254.69	Winter	FALSE	0.5525	five stars abs	TRUE
1	ORD493134	ID31309992!	13/1/2019	Bakers	[('Toshika 75(	23915	68.18	-37.81427	145.0111271	5	22787.43	Summer	TRUE	1.4766	not unlocked	FALSE
2	ORD039094	ID05914310:	21/2/2019	Nickolson	[('Universe N	24075	68.18	-37.817884	144.967848	0	24143.18	Summer	TRUE	0.1694	None	TRUE
3	ORD208432	ID11320536!	21/3/2019	Thompson	[('Toshika 75(	11015	63.54	-37.81791	144.946206	25	8324.79	Spring	FALSE	0.5879	three stars it'	TRUE
4	ORD271000	ID03736558(	12/5/2019	Thompson	[('Lucent 330	6055	70.79	-37.812655	144.9252436	25	4612.04	Autumn	FALSE	1.9194	absolutely lo	TRUE
5	ORD079617	ID38045691	16/11/2019	Bakers	[('Lucent 330	9630	105.32	-37.82023	144.989425	0	9735.32	Spring	FALSE	1.2484	very good qu	TRUE
6	ORD350274	ID01227525	25/11/2019	Thompson	[('Thunder lin	10335	100.65	-37.814123	144.9397096	0	10435.65	Spring	FALSE	0.667	five stars gre	TRUE
7	ORD325153	ID27172371	11/2/2019	Nickolson	[('Candle Infe	4880	107.41	-37.798829	144.9692883	15	4255.41	Summer	FALSE	2.2004	good product	TRUE
8	ORD060821	ID06347777	17/1/2019	Thompson	[('pearTV', 2),	28160	72.55	-37.814846	144.937541	15	24008.55	Summer	FALSE	0.8721	is this a stole	TRUE
9	ORD219114	ID22393897(	21/3/2019	Bakers	[('Alcon 10', 1	12400	62.01	-37.815402	144.999766	10	11222.01	Autumn	TRUE	1.3428	dead on arriv	FALSE
10	ORD219207	ID61674409!	8/3/2019	Nickolson	[('Candle Infe	7760	63.65	-37.816182	144.963919	25	5883.65	Autumn	FALSE	0.5634	make sure yo	TRUE
11	ORD075113	ID20208138:	13/3/2019	Thompson	[('Thunder lin	13350	81.52	-37.810675	144.958821	0	13431.52	Autumn	TRUE	1.0572	five stars the	TRUE
12	ORD311767	ID02504805!	20/3/2019	Bakers	[('Olivia x460	15540	64.38	145.0050339	-37.81177	0	15604.38	Autumn	FALSE	0.8844	great item pe	TRUE
13	ORD223250	ID06410259	30/11/2019	Thompson	[('iStream', 2)	37720	76.1	-37.809274	144.937567	15	32138.1	winter	FALSE	0.9174	good value o	TRUE
14	ORD085009	ID37750197!	18/5/2019	Bakers	[('Alcon 10', 2	21350	66.29	-37.807898	144.983595	25	15468.32	Autumn	FALSE	1.0498	fast delivery	TRUE
15	ORD112491	ID49097767	8/6/2019	nickolson	[('Universe N	14210	80.62	-37.821656	145.002561	15	12159.12	Winter	TRUE	1.4492	five stars love	TRUE
16	ORD139055	ID01371810	29/3/2019	nickolson	[('Lucent 330	12785	68.97	-37.801589	144.962194	10	11575.47	Autumn	FALSE	1.8144	value for mor	TRUE
17	ORD172436	ID05801719:	28/6/2019	Thompson	[('Candle Infe	7970	81.99	-37.80192	144.934318	5	7653.49	Winter	TRUE	1.6403	worth the ca	FALSE
18	ORD136137	ID07777835	8/4/2019	Thompson	[('Lucent 330	2760	63.21	-37.805423	144.953791	25	2133.21	Autumn	FALSE	1.0004	no home butt	FALSE
19	ORD376718	ID03560964	17/6/2019	Bakers	[('Alcon 10', 2	2610	78.26	-37.819002	144.991971	0	30928.26	Winter	TRUE	1.0427	best ever!!! v	TRUE
20	ORD462947	ID57977240	27/10/2019	Thompson	[('iStream', 1)	4825	99.72	-37.806999	144.945797	10	4442.22	Spring	TRUE	0.6415	nice tel√©fo	TRUE
21	ORD329313	ID07777834:	27/6/2019	Thompson	[('Universe N	5900	63.38	-37.804794	144.9533550	5	5668.38	Winter	FALSE	1.0368	great value w	TRUE
22	ORD070722	ID07117735	11/4/2019	Thompson	[('Universe N	8820	64.03	-37.817578	144.948678	5	8443.03	Autumn	FALSE	0.5641	four stars ver	TRUE
23	ORD182134	ID16808895!	23/3/2019	Bakers	[('pearTV', 2),	15575	65.17	-37.803277	145.000589	10	14082.67	Autumn	FALSE	0.884	awesome cel	TRUE
24	ORD487872	ID38077480!	13/7/2019	Bakers	[('iAssist Line	2525	75.04	-37.806785	144.998245	15	2221.29	Winter	TRUE	0.445	awesome ph	TRUE
25	ORD474479	ID01299536(	29/4/2019	Bakers	[('Universe N	9080	60.88	-37.809844	144.993821	25	6870.88	Autumn	FALSE	0.1252	g6 better pho	TRUE
26	ORD145274	ID03683895	15/5/2019	Nickolson	[('Lucent 330	4640	69.7	-37.801391	144.9706123	5	4477.7	Autumn	FALSE	1.9174	love it, we ar	TRUE
27	ORD402963	ID60213209	2019-23-04	Thompson	[('Thunder lin	4790	64.41	-37.816418	144.9382464	10	4375.41	Autumn	FALSE	0.8807	great conditi	TRUE
28	ORD180921	ID06502736	27/10/2019	Nickolson	[('Lucent 330	6295	80.09	-37.811372	144.98014	10	5745.59	Spring	FALSE	1.2302	the speaker,"	TRUE
29	ORD085779	ID07810051!	14/9/2019	Thompson	[('iStream', 2)	11790	78.18	-37.805425	144.952137	15	10099.68	summer	FALSE	0.9218	nice upgrade	TRUE
30	ORD037420	ID61671917	31/10/2019	Nickolson	[('Universe N	10880	76.61	-37.813358	144.974138	5	8617.11	Spring	FALSE	0.7089	best on mark	TRUE
31	ORD268549	ID05793930!	20/8/2019	Nickolson	[('Universe N	18250	50.82	-37.820308	144.958898	15	15563.32	Winter	FALSE	0.956	could not rec	TRUE
32	ORD102488	ID15601494	14/3/2019	Thompson	[('Universe N	19520	79.6	-37.816395	144.93817	0	19599.6	Autumn	TRUE	0.8855	easily the be	TRUE

For is_expidited we decided to leverage the linear dependence of the is_expidited column on delivery charge and built a model to decide at which threshold(for which rows) the is_expidited column should be updated, and used R squared to evaluate the accuracy.

2. Outlier Data

Feedback provided:

a. The model needs to be checked for accuracy after dropping the outliers.

Improvements after Feedback :

We dropped the records that were detected as outliers, and checked the accuracy of our model, through this we were able to alter our outlier detection metric , i.e we had use the 3 sigma rule and in order to decide a threshold for outliers we had attempted different values for the confidence interval , since when 3 sigma was giving us a very low accuracy , not all outliers had been dropped we attempted a 2 standard deviation and this gave us a better R squared value.

3. Task 2:

Feedback provided:

a. Research if the target variable needs to be transformed.

Improvements after Feedback:

On performing research through various online posts and papers we were able to draw a conclusion that target variables and not usually transformed. This is because we want to maintain the distribution and we want to avoid interpretability issues once predictions are made.