Big Data Analysis- a project diary

Keywords: SQL, MSSQL, Database, Importing data, Data acquisition, Kaggle, dataset, Exploratory Data Analysis (EDA)

Step 1 – Setting up MSSQL and restoring database

- Create a local database by downloading and installing MSSQL from Microsoft
    - Conclusion: Setting up a MSSQL database was easy with the help from online sources
    - Challenges: Having close to no knowledge in how to setup a database made it difficult
    - Cool Techniques used: Good YouTube video (https://www.youtube.com/watch?v=S2zBHmkRbhY)
    - Further development: Tinker with the admin side of the database

- Import the dataset

    The dataset is from Kaggle and is called "10Million Rows Turkish Market Sales Dataset(MSSQL)" (https://www.kaggle.com/datasets/omercolakoglu/10million-rows-turkish-market-sales-dataset) and is a MSSQL backup file containing 10M+ rows of data, with 27k supermarket items, 81 stores, 100k different customers, every order contains 1-9 different items and the numbers of each item ordered are 1-9.

    I restored the database by selecting Task -> Restore Database.

Step 2 – Initial Exploration of the dataset (understanding the data)

Looking at what columns we have

- Use Microsoft SQL Server Management Studio to look at the data
    - SELECT TOP 10 * FROM dbo.SALES; shows the first 10 records of the table

I can see that the are various columns such as:

- ID (datatype: int) – a general ID number
- ORDERID (int) – order number presumably ordered by order date
- ORDERDETAILID (int) – unknown. Further investigations needed
- DATE_ (datetime) - presumably the date of the order
- USERID (int) – presumably a unique customer id number
- USERNAME_ (char) – email of customer
- NAMESURNAME (char) – full name of customer
- STATUS (int) – unknown
- ITEMID (int) – unique ID of item ordered?
- ITEMCODE (int) – hmm…need to check this and ITEMID to see what they are
- ITEMNAME (char) – names of ordered item
- AMOUNT (int) – number ordered of a specific item

- UNITPRICE (float) – price of item
- PRICE (float) – unsure. Needs to be checked. Maybe price without VAT?
- TOTALPRICE (float) – seems to be amount * price at first glance, but needs to be checked in conjunction with PRICE
- CATEGORY1 (char) – first category of where item belongs to?
- CATEGORY2 (char) – first under-category of item?
- CATEGORY3 (char) - second under-category of item?
- CATEGORY4 (char) – third under-category of item?
- BRAND (char) – brand name of item
- USERGENDER (char) – gender of customer
- USERBIRTHDATE (datetime) – birthdate of customer
- REGION (char) – region where the store is located?
- CITY (char) – city where store is located?
- TOWN (char) – suburb or area of city?
- DISTRICT (char) – district of city/town?
- ADDRESTEXT (char) - full address of store?
- ADDRESSID (NULL) – no values?

So, we have a lot of exploring to do to understand the dataset. The dataset is 10M+ orders placed in Türkiye (Turkey) and the names and genders are in Turkish and need to be translated. Furthermore, we need to investigate the dataset to understand the columns and how they relate to each other.

Exploring the columns to understand how they relate to one another

I decided to take explore the columns to see if they have unique values or not, and if so – how many

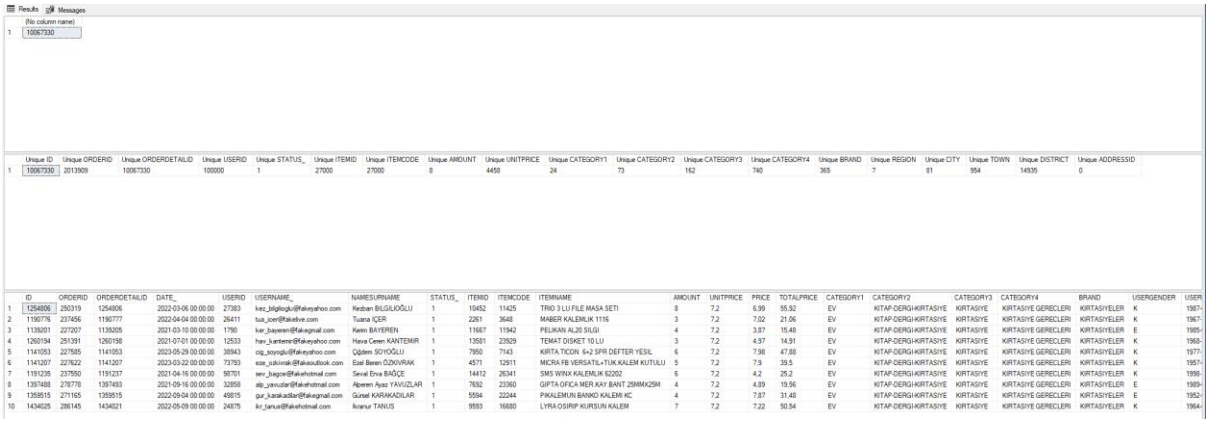
I queried:

```
SELECT COUNT(*)
FROM dbo.SALES;

SELECT
       COUNT(DISTINCT ID) as "Unique ID"
       ,COUNT(DISTINCT ORDERID) as "Unique ORDERID"
       ,COUNT(DISTINCT ORDERDETAILID) as "Unique ORDERDETAILID"
       ,COUNT(DISTINCT USERID) as "Unique USERID"
       ,COUNT(DISTINCT STATUS_) as "Unique STATUS_"
       ,COUNT(DISTINCT ITEMID) as "Unique ITEMID"
       ,COUNT(DISTINCT ITEMCODE) as "Unique ITEMCODE"
       ,COUNT(DISTINCT AMOUNT) as "Unique AMOUNT"
       ,COUNT(DISTINCT UNITPRICE) as "Unique UNITPRICE"
       ,COUNT(DISTINCT CATEGORY1) as "Unique CATEGORY1"
       ,COUNT(DISTINCT CATEGORY2) as "Unique CATEGORY2"
       ,COUNT(DISTINCT CATEGORY3) as "Unique CATEGORY3"
       ,COUNT(DISTINCT CATEGORY4) as "Unique CATEGORY4"
       ,COUNT(DISTINCT BRAND) as "Unique BRAND"
       ,COUNT(DISTINCT REGION) as "Unique REGION"
       ,COUNT(DISTINCT CITY) as "Unique CITY"
       ,COUNT(DISTINCT TOWN) as "Unique TOWN"
       ,COUNT(DISTINCT DISTRICT) as "Unique DISTRICT"
```

```
        ,COUNT(DISTINCT ADDRESSID) as "Unique ADDRESSID"
FROM
        dbo.SALES;

SELECT
        TOP 10 *
FROM
        dbo.SALES;
```



From this, we can see that:

- number of rows are 10 067 330.
- ID is just a unique id for each row, as expected.
- We seem to have 2M+ orders (ORDERID)
- The ORDERDETAIL and ID seems to be identical (need to verify!)
- USERID – we seem to have 100 000 unique customers
- STATUS_ seems to be only 1 (as in ordered confirmed?), thus lacks value for us
- ITEMID and ITEMCODE are both 27k, but not identical (more exploration needed!)
- AMOUNT seems to have values 1, 2, 3, 4, 5, 6, 7, 8 or 9 (as per description, verification needed!)
- UNITPRICE have 4458 different prices
- CATEGORY1 have 24 unique values
- CATEGORY2 have 73 unique values
- CATEGORY3 have 162 unique values
- CATEGORY4 have 740 unique values (these values indicates that CATEGORY1 is a Tier1 category, and that the others are under-categories to each other (need verification!)
- BRAND – we have 365 unique brands
- REGION – we have 7 country regions
- CITY – we have 81 cities
- TOWN – we have 954 towns (smaller settlement not as big as a city? Needs to be verified!)
- DISTRICT – we have 14 935 unique districts belonging to different parts of a city/town
- ADDRESSID have no values

Great information!
We've now gotten a better picture of the data and how they are related.

But we still need to verify some things like:

1. Are the ORDERDETAILID and ID ever different? If not, then ORDERDETAILID is redundant.
2. Make sure that STATUS only have values of "1". If so, it is redundant information.
3. Why are ITEMID and ITEMCODE not identical, but they have the same amount of unique number of values (27 000)
4. IS AMOUNT ever any other numbers but 1 through 9?
5. Look at text of the categories and translate them in order to understand them in relation to each other. Later, we also need to exchange the names in Turkish with names in English.
6. Verify that the REGIONSs are indeed Turkish country regions
7. Verify that the towns are indeed smaller settlements, than cities
8. Verify that DISTRICTS are districts within a city or town
9. Verify that the ADDRESSID is indeed NULL and thus redundant

1. Are the ORDERDETAINID and ID ever different?

```
-- Where do columns have identical values
SELECT
        COUNT(ID)
FROM
        dbo.SALES
WHERE
        ID=ORDERDETAILID
ORDER BY
        COUNT(ID)

SELECT
        ID, ORDERDETAILID
FROM
        dbo.SALES
WHERE ID=ORDERDETAILID


-- Where do columns NOT have idnetical values
SELECT
        COUNT(ID)
FROM
        dbo.SALES
WHERE
        ID<>ORDERDETAILID
ORDER BY
        COUNT(ID)

SELECT
```

```
        ID, ORDERDETAILID
FROM
        dbo.SALES
WHERE ID<>ORDERDETAILID
```

Give us:

	(No column name)
1	2014680

	ID	ORDERDETAILID
1	1254806	1254806
2	1141053	1141053
3	1141207	1141207
4	1359515	1359515
5	2129126	2129126
6	2242784	2242784
7	2625737	2625737
8	2738805	2738805
9	2912986	2912986
10	3297044	3297044
11	3988816	3988816
12	4090213	4090213
13	4530092	4530092
14	4954304	4954304
15	5605202	5605202
16	6628887	6628887
17	7493956	7493956
18	7856189	7856189

	(No column name)
1	8052650

	ID	ORDERDETAILID
1	1190776	1190777
2	1139201	1139205
3	1260194	1260198
4	1191235	1191237
5	1397488	1397493
6	1434025	1434021
7	1416223	1416224
8	1610551	1610549

We can see that about 2M+ of the 10M+ rows have identical values of ID and ORDERDETAILID, but that the rest, about 8M+ records do not have identical values. One can however see that they are usually not far off. Here it would be interesting to sort by ID to see what happens to ORDERDETAILID. Let's do that!

```
SELECT TOP 30 *
```

```
FROM dbo.SALES
ORDER BY ID
```

Gives:

	ID	ORDERID	ORDERDETAILID	DATE_	USERID	USERNAME_	NAMESURNAME	STATUS_	ITEMID	ITEMCODE	ITEMNAME	AMOUNT	UNITPRICE	PRICE	TOTALPRICE	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	BRAND
1	1	1	1	2022-02-20 00:00:00	20743	ali_dindaroglu@hotmail.com	Alye DINDAROĞLU	1	26548	41599	BIFA 204 KAK. KRM SAND.BIS.3LU 210 GR*10"	2	5.52	5.2	10.4	GIDA	BISKUVI-CEREZ	BISKUVI	KREMALI BISKUVI	BIFA
2	2	2	2	2021-11-21 00:00:00	63845	aya_agirtas@fakeyahoo.com	Ayşe AĞIRTAŞ	1	18955	33434	MAGGI COR YAYLA 70 GR *144*	8	4.32	3.2	25.6	GIDA	CORBA-BULYON	CORBALAR	HAZIR CORBA	MAGGI
3	3	3	9	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	5606	22266	COLGATE D.FIR. MAS.TOTAL PRO.GUM.SOFT *12*	7	51	45	315	KOZMETIK	AGIZ BAKIM	DIS FIRCALARI-IPLER	FIRCALAR	COLGATE
4	4	3	7	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	7830	6950	PALMIYE MELISA YAGI	2	16.7	15.95	31.9	KOZMETIK	KISISEL BAKIM	CILT BAKIM	YAGLAR	PALMIYE
5	5	3	8	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	4020	15947	ICIM SUT 200 ML CILEKLI *27*	8	8	7.13	57.04	KAHVALTILIK	SUT-YOGURT-PEYNIR	SUT	MEYVELI	ICIM
6	6	3	4	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	12228	5559	ULKER 034-II HERO BABY BEBE 800 GR TNK*4*	6	39.6	36.24	217.44	BEBEK	HAZIR YEMEK-MAMA	HAZIR COCUK YEMEKLERI	BEBE BISKUVISI	ULKER
7	7	3	6	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	6473	25656	TARIHE YON VEREN ZENGINLER VE YONETICILER	2	24.5	23.65	47.3	EV	KITAP-DERGI-KIRTASIYE	KITAP	KITAPLAR	KITAPLAR
8	8	3	3	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	22652	35047	KUTULU AYAKLI MIKROFON SETLI GITAR	2	322.25	287.13	574.26	OYUNCAK	ZEKA GELISTIRICI	OYUNCAKLAR	BEBE OYUNCAK	OYUNCAK
9	9	3	5	2022-03-21 00:00:00	33206	sey_turfanda@fakehotmail.com	Şeyda TURFANDA	1	23218	45152	TEMAT KOMBINE SET 2050 SIYAH	1	6	5.43	5.43	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER
10	10	4	12	2021-08-02 00:00:00	69767	yas_mestanlar@fakegmail.com	Yasin MESTANLAR	1	7275	13729	NERGIS YUFKA 1 KG	8	18	12.06	96.48	KAHVALTILIK	UNLU MAMULLER	YUFKA	BOREKLIK	NERGIS
11	11	4	11	2021-08-02 00:00:00	69767	yas_mestanlar@fakegmail.com	Yasin MESTANLAR	1	24112	36003	SNOPY SN-218 SIYAH-MIKROFONLU KULAKLIK	2	30	19.06	38.12	EV	ELEKTRIK-ELEKTRONIK	ELEKTRONIK	ELEKTRONIK ALETLER	SNOPY
12	12	4	10	2021-08-02 00:00:00	69767	yas_mestanlar@fakegmail.com	Yasin MESTANLAR	1	21959	20788	OYUNCAK BARBIE EV CAY SETI	8	317.4	208.41	1667.28	OYUNCAK	ZEKA GELISTIRICI	OYUNCAKLAR	BEBE OYUNCAK	OYUNCAK
13	13	5	15	2021-03-23 00:00:00	2692	rav_anthoom@fakehotmail.com	Ravza ANTHOOM	1	700	1696	NIVEA KREM SOFT 100 ML KAVANOZ	2	68.7	39.27	78.54	KOZMETIK	VUCUT-EL BAKIM	KREMLER	VUCUT-EL KREMLERI	NIVEA
14	14	5	14	2021-03-23 00:00:00	2692	rav_anthoom@fakehotmail.com	Ravza ANTHOOM	1	12670	22000	TOYBOX JELLY 80 GR MEY.BAHCESI *24*	3	9.4	5.75	17.25	SEKERLEME	SAKIZ-SEKERLEME	SEKERLEMELER	YUMUSAK SEKER	TOYBOX
15	15	5	13	2021-03-23 00:00:00	2692	rav_anthoom@fakehotmail.com	Ravza ANTHOOM	1	13223	14314	JIBER 130 ERK.PENYE BOXER	6	23.6	14.05	84.3	EV	TEKSTIL-GIYIM-AKSESUAR	CAMASIRLAR	ERKEK CAMASIR	JIBER
16	16	5	16	2021-03-23 00:00:00	2692	rav_anthoom@fakehotmail.com	Ravza ANTHOOM	1	22166	21078	TAT SALCA 710 GR CAM *12*	1	73.5	39.08	39.08	GIDA	HAZIR YEMEK-KONSERVE-SALCA	SALCA	DOMATES	TAT
17	17	6	18	2023-05-16 00:00:00	9604	gul_celikolu@fakeyahoo.com	Gülten CELIKOLU	1	22408	34624	DANONE P.DANINO 6X40 GR MUZ	4	13.56	15.38	61.52	SUT	DONDURMA-SUTLU TATLI	SUTLU TATLILAR	MEYVELI	DANONE
18	18	6	19	2023-05-16 00:00:00	9604	gul_celikolu@fakeyahoo.com	Gülten CELIKOLU	1	13964	29041	COLGATE TOTAL GEL FERAHLIK 50 ML*12*	8	95.1	110.14	881.12	KOZMETIK	AGIZ BAKIM	DIS MACUNLARI	MACUNLAR	COLGATE
19	19	6	17	2023-05-16 00:00:00	9604	gul_celikolu@fakeyahoo.com	Gülten CELIKOLU	1	19880	31182	TILLO 1656 LUX ASKI 6'LI *24*	4	45.3	48.88	195.52	TEMIZLIK	EV GERECLERI	EV PLASTIK GERECLERI	PLASTIK	TILLO
20	20	6	20	2023-05-16 00:00:00	9604	gul_celikolu@fakeyahoo.com	Gülten CELIKOLU	1	25797	36848	GEZER FILET BABET KIZ GAZYFM 01883.00	2	97	112.6	225.2	EV	TEKSTIL-GIYIM-AKSESUAR	AYAKKABI-TERLIK	AYAK GIYIM	GEZER
21	21	7	23	2023-06-08 00:00:00	13326	sel_baspinar@fakeoutlook.com	Selahattin BAŞPINAR	1	271	693	KARLIDAG PARMAK PEYNIRI KG	5	60.55	70.99	425.94	KAHVALTILIK	SUT-YOGURT-PEYNIR	PEYNIRLER	PARMAK PEYNIRLER	KARLIDAG
22	22	7	22	2023-06-08 00:00:00	13326	sel_baspinar@fakeoutlook.com	Selahattin BAŞPINAR	1	16650	18931	MABER TAHTA KALEMI KARTUSLU SIYAH	5	17.95	20.2	101	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER
23	23	7	24	2023-06-08 00:00:00	13326	sel_baspinar@fakeoutlook.com	Selahattin BAŞPINAR	1	23283	39542	IGLO KREMALI MINI BOREK 500 GR *10*	2	60	69.28	138.56	GIDA	DONUK URUNLER	DONDURULMUS GIDA	BOREK	IGLO
24	24	7	21	2023-06-08 00:00:00	13326	sel_baspinar@fakeoutlook.com	Selahattin BAŞPINAR	1	24263	35568	KIRTA TEMAT STICK YAPIS 40GR	5	28.75	31.46	157.3	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER
25	25	8	25	2022-01-31 00:00:00	13352	rav_kapi@fakeoutlook.com	Ravza KAPI	1	18368	30946	CARS BUYUK POTA	2	297.25	273.24	546.48	OYUNCAK	ZEKA GELISTIRICI	OYUNCAKLAR	BEBE OYUNCAK	OYUNCAK
26	26	9	27	2022-05-09 00:00:00	32622	mah_ozarin@fakegmail.com	Mahmut ÖZARIN	1	3439	8898	SUN SC-420 MINI KEK KAPSULLU (90LI)	5	29.5	29.66	148.3	EV	MUTFAK GERECLERI	MUTFAK ESYA-GERECLERI	PLASTIK ALTLIKLAR	SUN
27	27	9	26	2022-05-09 00:00:00	32622	mah_ozarin@fakegmail.com	Mahmut ÖZARIN	1	20017	31374	EMOTION EDT+DEO KDFRE OCEAN FRESH*6*	6	274.95	258.59	1551.54	KOZMETIK	PARFUM-DEODORANT	DEODORANTLAR	DEO	EMOTION
28	28	9	28	2022-05-09 00:00:00	32622	mah_ozarin@fakegmail.com	Mahmut ÖZARIN	1	20694	34060	DERGI MERAKLI MINIK	1	30	30.32	30.32	EV	KITAP-DERGI-KIRTASIYE	DERGI	HAFTALIK	DERGILER
29	29	10	33	2021-01-18 00:00:00	95398	cen_canikare@fakeoutlook.c...	Cennet Ebrar CANI...	1	7906	7065	CILEKKIZ RULO KAP	6	8.05	4.48	26.88	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER
30	30	10	30	2021-01-18 00:00:00	95398	cen_canikare@fakeoutlook.c...	Cennet Ebrar CANI...	1	2499	8201	PEYMANCIFTEKAV.FINDIK170GR*12*	3	80.45	46.8	140.4	GIDA	BISKUVI-CEREZ	KURUYEMIS	DIGER KURUYEMIS	PEYMAN

Ok. We can see that for each ORDERID we have a unique ORDERDETAILID. Each ORDERID is assigned a unique ORDERDETAILID. We can see that ORDERID=1 and ORDERID=2 both only contain one item ordered, while ORDERID=3 have 7 different items ordered within the same order. Since ORDERDETAILID=1 and ORDERDETAILID=2 already have been assigned to ORDERID=1 and ORDERID=2, we now get to assign ORDERDETAILID=3 through ORDERDETAILID=9 (7 different items ordered). It's a bit weird that the ORDERDETAILID is not grouped as ORDERDETAILID=3 through ORDERDETAILID=9, but now we at least understand that ORDERDETAILID is a unique number based on ORDERID and the numbers of different items ordered. ORDERDETAILID seems to be redundant.

2. Make sure that STATUS only have values of "1"

Writing:

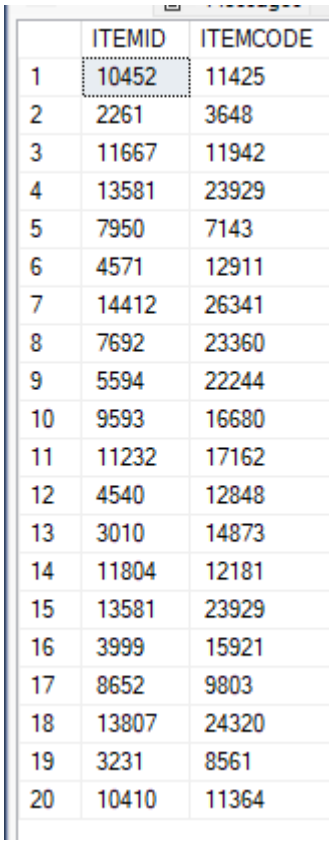
```
SELECT *
FROM dbo.SALES
WHERE STATUS_<>1;
```

Gives no rows. Thus STATUS_ is always 1, as in confirmed order or similar. Thus, it is redundant information and can be removed.

3. Why are ITEMID and ITEMCODE not identical, but they have the same amount of unique number of values (27 000)?

```
SELECT TOP 20 ITEMID, ITEMCODE
FROM dbo.SALES
```

Gives:



	ITEMID	ITEMCODE
1	10452	11425
2	2261	3648
3	11667	11942
4	13581	23929
5	7950	7143
6	4571	12911
7	14412	26341
8	7692	23360
9	5594	22244
10	9593	16680
11	11232	17162
12	4540	12848
13	3010	14873
14	11804	12181
15	13581	23929
16	3999	15921
17	8652	9803
18	13807	24320
19	3231	8561
20	10410	11364

Looks like we have different values indeed. Let's see if there are ever any rows with identical values in ITEMID and ITEMCODE, then we have to order by ITEMID and then ITEMCODE to see what we can learn from that.

```
SELECT COUNT(*)
FROM dbo.SALES
WHERE ITEMID=ITEMCODE
```

Gives: 2607

So, we have very few (coincidence?) values where ITEMID and ITEMCODE are the same.

Displaying these via:

```sql
SELECT COUNT(*)
FROM dbo.SALES
WHERE ITEMID=ITEMCODE

SELECT *
FROM dbo.SALES
WHERE ITEMID=ITEMCODE
```

Gives:

	(No column name)
1	2607

	ID	ORDERID	ORDERDETAILID	DATE_	USERID	USERNAME_	NAMESURNAME	STATUS_	ITEMID	ITEMCODE	ITEMNAME	AMOUNT	UNITPRICE	PRICE	TOTALPRICE	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	BRAND	USERGENDER	USE
1	1875743	374385	1875742	2021-07-07 00:00:00	75199	yak_sokuk@fakeoutlook.com	Yakup SÖKÜK	1	20346	20346	WINKS TENEKE KALEM TRAS	6	7	4,69	28,14	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	198
2	2242217	447826	2242217	2021-03-17 00:00:00	24982	eli_haliloglu@fakelive.com	Elfsu HALILOGLU	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	3	6,05	3,46	10,38	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	195
3	6051254	1210198	6051254	2022-03-16 00:00:00	90487	ebr_camasiroglu@fakegmail.com	Ebru Şeyma ÇAMAŞIROĞLU	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	2	6,05	5,71	11,42	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	197
4	7072746	1414827	7072745	2023-02-14 00:00:00	82027	gul_cakmaktas@fakeyahoo.com	Gülay ÇAKMAKTAŞ	1	20346	20346	WINKS TENEKE KALEM TRAS	7	7	7,41	51,87	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	199
5	9569933	1914579	9569929	2022-11-28 00:00:00	77043	nur_diri@fakeyahoo.com	Nuray DIRI	1	20347	20347	WINKS OGRENCI MAKASI	1	5,6	5,94	5,94	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	196
6	4222782	844234	4222780	2021-06-30 00:00:00	39024	alp_resiti@fakeyahoo.com	Alper REŞİTİ	1	20346	20346	WINKS TENEKE KALEM TRAS	5	7	4,42	22,1	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	197
7	5317405	1063312	5317403	2022-02-13 00:00:00	20768	zum_begici@fakelive.com	Zümra BEGICI	1	20347	20347	WINKS OGRENCI MAKASI	5	5,6	5,1	25,5	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	199
8	8913682	1783250	8913682	2021-02-04 00:00:00	87783	afi_zerener@fakeyahoo.com	Afife ZERENER	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	6	6,05	3,42	20,52	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	195
9	5370767	1073897	5370771	2023-05-31 00:00:00	42469	ara_camdeviren@fakeoutlook.com	Aras ÇAMDEVIREN	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	8	6,05	6,74	53,92	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	196
10	6411137	1282261	6411134	2021-03-17 00:00:00	64184	ebu_isbeceren@fakegmail.com	Ebubekir IŞBECEREN	1	20346	20346	WINKS TENEKE KALEM TRAS	4	7	4,13	16,52	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	198
11	6412906	1282619	6412910	2022-04-11 00:00:00	54878	sel_kokboru@fakehotmail.com	Selma KÖKBORU	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	6	6,05	5,49	32,94	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	196
12	7354860	1471302	7354860	2021-01-24 00:00:00	83810	ecr_siranturkekul@fakehotmail.com	Ecrin ŞIRANTÜRKEKUL	1	20347	20347	WINKS OGRENCI MAKASI	5	5,6	3,23	16,15	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	196
13	6773420	1354865	6773414	2021-08-04 00:00:00	80836	seb_belgi@fakegmail.com	Sebahattin BELGI	1	20346	20346	WINKS TENEKE KALEM TRAS	6	7	4,6	27,6	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	198
14	7913808	1583225	7913805	2021-04-16 00:00:00	28022	hal_uysan@fakegmail.com	Halil Yunus UYSAN	1	20347	20347	WINKS OGRENCI MAKASI	2	5,6	3,54	7,08	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	199
15	1877542	374761	1877543	2022-11-26 00:00:00	41241	mer_ugus@fakeyahoo.com	Meryem Ezgi UĞUŞ	1	20347	20347	WINKS OGRENCI MAKASI	2	5,6	5,89	11,78	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	199
16	5541236	1108097	5541237	2023-06-09 00:00:00	75953	ker_erguzer@fakeoutlook.com	Kerim ERGÜZER	1	6870	6870	KIRTA.MET.80 YAP.KARELI METOD	5	6,05	7,02	35,1	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	196
17	7087032	1417629	7087034	2023-04-07 00:00:00	5685	esm_tosunbayrak@fakeyahoo.com	Esma TOSUNBAYRAK	1	20347	20347	WINKS OGRENCI MAKASI	7	5,6	6,13	42,91	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	195

We can see that certain ITEMIDs show up several times where they are identical to ITEMCASE. Let's see if ITEMID=6870 can have other ITEMCODEs than 6870.

```sql
SELECT *
FROM dbo.SALES
WHERE (ITEMID=6870 AND ITEMID<>ITEMCODE)
```

Gives: No rows. Thus, some ITEMIDs and ITEMCODEs are the same.

Let's sort by ITEMID and see what we can learn.

```sql
SELECT TOP 1500 *
FROM dbo.SALES
ORDER BY ITEMID ASC
```

Gives (selected view where there's a change in values of ITEMID:

368	5837106	1167262	5837112	2023-04-09 00:00:00	5658	nai_eyyuboglu@fakeyahoo.com	Naime EYYÜBOĞLU	1	1	5	PIL KODAK XTRA HEAVY 9 V	4	25,75	27,8	111,2	EV	ELEKTRIK-ELEKTRONIK	PIL	PIL	KODAK	K	1995-04-18	Do
369	8015522	1603614	8015528	2022-11-08 00:00:00	10676	raz_dokmeoglu@fakegmail.com	Raziye Tansu DÖKMEOĞ...	1	1	5	PIL KODAK XTRA HEAVY 9 V	1	25,75	28,98	28,98	EV	ELEKTRIK-ELEKTRONIK	PIL	PIL	KODAK	K	1950-02-25	Ma
370	6259332	1251932	6259336	2022-10-12 00:00:00	18672	ken_topcuoglu@fakegmail.com	Kenan Diyar TOPCUOGLU	1	1	5	PIL KODAK XTRA HEAVY 9 V	1	25,75	26,79	26,79	EV	ELEKTRIK-ELEKTRONIK	PIL	PIL	KODAK	E	1952-05-29	Eg
371	3658793	731416	3658796	2022-12-22 00:00:00	31066	abd_sargan@fakeoutlook.com	Abdulsamet Bilal SARGAN	1	1	5	PIL KODAK XTRA HEAVY 9 V	2	25,75	27,92	55,84	EV	ELEKTRIK-ELEKTRONIK	PIL	PIL	KODAK	E	1978-10-23	Ma
372	5955395	1191017	5955400	2021-09-01 00:00:00	33420	ays_dalcintutan@fakehotmail.com	Ayşegül DALÇINTUTAN	1	2	6	PIL KODAK AA*2 MAX ALK...	4	41,3	28,18	112,72	EV	ELEKTRIK-ELEKTRONIK	PIL	KALEM PILLER	KODAK	K	1961-02-12	Eg
373	1051063	209677	1051064	2021-09-24 00:00:00	28635	erk_kitir@fakegmail.com	Erkan KITIR	1	2	6	PIL KODAK AA*2 MAX ALK...	5	41,3	30,56	152,8	EV	ELEKTRIK-ELEKTRONIK	PIL	KALEM PILLER	KODAK	E	1973-01-19	İç
374	9450010	1890420	9450010	2021-08-13 00:00:00	71328	ser_karakayaci@fakegmail.com	Şerife KARAKAYACI	1	2	6	PIL KODAK AA*2 MAX ALK...	8	41,3	28,88	231,04	EV	ELEKTRIK-ELEKTRONIK	PIL	KALEM PILLER	KODAK	K	1980-06-20	Eg
375	1619806	323274	1619813	2022-07-02 00:00:00	98748	mer_unalan@fakelive.com	Meryem Döndü ÜNALAN	1	2	6	PIL KODAK AA*2 MAX ALK...	8	41,3	41,91	335,28	EV	ELEKTRIK-ELEKTRONIK	PIL	KALEM PILLER	KODAK	K	1970-03-19	Ka

We can see that CATEGORY 4 as well as ITEMNAME changes.

Running:

```
SELECT *
FROM dbo.SALES
WHERE ITEMID IN (200)
ORDER BY ITEMID ASC
```

And changing 200 with 1 and 2 etc, I can see that ITEMCODE is the same for each unique ITEMID.

Thus, ITEMCODE seems redundant.

4. IS AMOUNT ever any other numbers but 1 through 9?

```
SELECT *
FROM dbo.SALES
WHERE AMOUNT NOT IN (1, 2, 3, 4, 5, 6, 7, 8, 9)
```

Gives: No rows. Thus, AMOUNT only have values between 1 and 9.

We will probably be interested in answering business questions such as what categories sells most/worst, but also what items. So, my thought is to translate the categories and leave the item names "as is". Maybe identify the top 5 items when that time comes, but not now.

Before translating and changing names of categories, I think it is time to create a new view or table where we drop certain redundant columns identified recently.

Columns identified as redundant are:

- ORDERDETAILID
- USERNAME_
- STATUS_
- ITEMCODE
- ADDRESSID

```
SELECT DISTINCT CATEGORY1
FROM dbo.SALES
ORDER BY CATEGORY1 ASC
```

Gives:

	CATEGORY1
1	BALIK
2	BEBEK
3	CAY-KAHVE-SEKER
4	DETERJAN
5	ET
6	EV
7	GIDA
8	KAGIT
9	KAHVALTILIK
10	KARO
11	KOZMETIK
12	KUMES
13	MEYVE
14	MUHTELIF
15	OYUNCAK
16	SARF
17	SEBZE
18	SEKERLEME
19	SICAK ICECEKLER
20	SIGARALAR
21	SOGUK ICECEKLER
22	SUT
23	TEMIZLIK
24	YESILLIK

Step 3 – Creating a new table without redundant columns

In order to preserve the old database and have a working table with which I can make changes to, I will create a new table from an existing one.

The columns we have are generated from this query:

```
SELECT COLUMN_NAME
FROM INFORMATION_SCHEMA.COLUMNS
WHERE TABLE_NAME='SALES'
```

Gives:

	COLUMN_NAME
1	ID
2	ORDERID
3	ORDERDETAILID
4	DATE_
5	USERID
6	USERNAME_
7	NAMESURNAME
8	STATUS_
9	ITEMID
10	ITEMCODE
11	ITEMNAME
12	AMOUNT
13	UNITPRICE
14	PRICE
15	TOTALPRICE
16	CATEGORY1
17	CATEGORY2
18	CATEGORY3
19	CATEGORY4
20	BRAND
21	USERGENDER
22	USERBIRTHDATE
23	REGION
24	CITY
25	TOWN
26	DISTRICT
27	ADDRESSTEXT
28	ADDRESSID

We will only include the columns that we want, and exclude the ones that, above, were deemed redundant.

We use the following query:

```
SELECT
        ID
        ,ORDERID
        ,DATE_
        ,USERID
        ,NAMESURNAME
        ,ITEMID
        ,ITEMNAME
        ,AMOUNT
        ,UNITPRICE
        ,PRICE
        ,TOTALPRICE
        ,CATEGORY1
        ,CATEGORY2
        ,CATEGORY3
        ,CATEGORY4
        ,BRAND
        ,USERGENDER
```

```
        ,USERBIRTHDATE
        ,REGION
        ,CITY
        ,TOWN
        ,DISTRICT
        ,ADDRESSTEXT
INTO
        DATA
FROM
        SALES
```

Gives: A new table with which we can change names without affecting the original database table, ‘SALES’. The new table ‘DATA’ has no redundant data.

Querying (to check that all redundant columns are gone and the table has been created):

```
SELECT TOP 10 *
FROM DATA
```

Gives:

	ID	ORDERID	DATE_	USERID	NAMESURNAME	ITEMID	ITEMNAME	AMOUNT	UNITPRICE	PRICE	TOTALPRICE	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	BRAND	USERGENDER	USERBIRTHDATE	REGION	CITY	TOWN	DISTRICT	ADDRESSTEXT
1	1254806	250319	2022-03-06 00:00:00	27383	Kedsen BILGILIOĞLU	10452	TRIO 3 LU FILE MASA SETI	8	7.2	6.99	55.92	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1967-09-17	Marmara	Istanbul	BÜYÜKÇEKMECE	MURATBEY MERKEZ MAH.	MURATBEY MERKEZ MAH. 96. SOI
2	1190776	237456	2022-04-04 00:00:00	26411	Tuana İÇER	2261	MABER KALEMLIK 1116	3	7.2	7.02	21.06	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1967-12-28	Marmara	Istanbul	BÜYÜKÇEKMECE	MURAT ÇEŞME MAH.	MURAT ÇEŞME MAH. ILAYDA SOK
3	1138201	227207	2021-03-10 00:00:00	1790	Kerim BAYEREN	11667	PELIKAN ALÇE SILGI	4	7.2	3.87	15.48	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	1989-09-02	Marmara	Istanbul	BÜYÜKÇEKMECE	CUMHURIYET MAH.	CUMHURIYET MAH. IBRAHIM ERK
4	1260194	251391	2021-07-01 00:00:00	12533	Hava Ceren KANTEMİR	13581	TEMAT DISKET 10 LU	3	7.2	4.97	14.91	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1968-11-20	Marmara	Istanbul	BÜYÜKÇEKMECE	GÜZELCE MAH.	GÜZELCE MAH. ISIMSIZ_153 SOKA
5	1141053	227585	2023-05-29 00:00:00	38943	Çiğdem SOYOĞLU	7950	KIRTA TICON 6+2 SPR DEFTER YESIL	6	7.2	7.98	47.88	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1977-10-05	Marmara	Istanbul	BÜYÜKÇEKMECE	KAMILOBA MAH.	KAMILOBA MAH. BEYKENT SOKAK
6	1141207	227622	2023-03-22 00:00:00	73793	Esel Beren DÖNÜRAK	4571	MICRA FB VERSATIL+TUK KALEM KUTULU	5	7.2	7.9	39.5	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1957-07-14	Marmara	Istanbul	BÜYÜKÇEKMECE	KAMILOBA MAH.	KAMILOBA MAH. GEZGIN SOKAK
7	1191235	237550	2021-04-16 00:00:00	98701	Seval Erva BAĞÇE	14412	SMS WINK KALEMLIK 62202	6	7.2	4.2	25.2	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1996-12-17	Marmara	Istanbul	BÜYÜKÇEKMECE	KUMBURGAZ MERKEZ MAH.	KUMBURGAZ MERKEZ MAH. ISIMS
8	1397468	276778	2021-09-16 00:00:00	32858	Alperen Ayaz YAVUZLAR	7692	GIPTA OFICA MER.KAY BANT 25MMX25M	4	7.2	4.89	19.56	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	1989-01-04	Marmara	Istanbul	BÜYÜKÇEKMECE	KUMBURGAZ MERKEZ MAH.	KUMBURGAZ MERKEZ MAH. SILIV
9	1359515	271165	2022-09-04 00:00:00	49815	Günsel KARAKADILAR	5594	PIKALEMUN BANKO KALEM 4C	4	7.2	7.87	31.48	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	E	1952-06-13	Marmara	Istanbul	BÜYÜKÇEKMECE	CELALIYE MAH.	CELALIYE MAH. GÖNÜLDEN SOKA
10	1434625	286145	2022-05-09 00:00:00	24875	İnanur TANIS	9593	LYRA OSIRIP KURSUN KALEM	7	7.2	7.22	50.54	EV	KITAP-DERGI-KIRTASIYE	KIRTASIYE	KIRTASIYE GERECLERI	KIRTASIYELER	K	1964-08-09	Marmara	Istanbul	BÜYÜKÇEKMECE	MURAT ÇEŞME MAH.	MURAT ÇEŞME MAH. 324. SOKAK

So, a success!

5. Translate categories to English

Now let's do some changes to the table by translating some words.

By running the query above to get the distinct categories of 'CATAGORY1'. These are copied to excel and then I use Google Translate to translate the names. With some smaller changes I got this:

BALIK	FISH
BEBEK	BABY
CAY-KAHVE-SEKER	TEA-COFFEE-SUGAR
DETERJAN	DETERGENT
ET	MEAT
EV	HOUSE
GIDA	FOOD
KAGIT	PAPER
KAHVALTILIK	FOR BREAKFAST
KARO	TILES
KOZMETIK	COSMETIC
KUMES	FABRICS
MEYVE	FRUIT
MUHTELIF	MISCELLANEOUS
OYUNCAK	TOY
SARF	CONSUMPTION
SEBZE	VEGETABLES
SEKERLEME	CONFECTIONERY
SICAK ICECEKLER	HOT DRINKS
SIGARALAR	CIGARETTES
SOGUK ICECEKLER	COLD DRINKS
SUT	MILK
TEMIZLIK	CLEANING SUPPLIES
YESILLIK	GREENS

To make sure that the category translation is correct, I check the top 10 rows of data of each category and then take the ITEMNAME and run it through Google Images. Like:

```
SELECT TOP 10 *
FROM DATA
WHERE CATEGORY1='BALIK'
```

Gives:

	ID	ORDERID	DATE_	USERID	NAMESURNAME	ITEMID	ITEMNAME	AMOUNT	UNITPRICE	PRICE	TOTALPRICE	CATEGORY1	CATEGORY2	CATEGORY3	CATEGORY4	BRAND
1	1154746	230302	2021-11-10 00:00:00	49195	Şahin EYİLER	19766	DARD TON LIGHT 160X2 GR "12"	1	42,9	30,17	30,17	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
2	1230344	245372	2023-01-01 00:00:00	75608	Hamza Duran FİDANCAN	19766	DARD TON LIGHT 160X2 GR "12"	8	42,9	44,37	354,96	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
3	1466492	292664	2021-01-21 00:00:00	15808	Emir ŞİLGİN	25539	DARD TON 185 X 2 GR FASULYELI "12"	2	39,66	19,7	39,4	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
4	1361920	271641	2021-04-24 00:00:00	74853	Muhammet Yiğit GÜREŞÇİ	12393	DARD.TON 3X80 GR HAY.BAL.YI. "16"	6	89,4	51,71	310,26	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
5	1573461	313993	2022-03-20 00:00:00	54786	Erdoğan Engin KUZUBAŞ	19764	DARD TON A.YAGLI 160 GR "24"	4	23,64	21,64	86,56	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
6	1469053	293174	2021-10-09 00:00:00	3941	Hamdi ZABUN	2760	DARD.TON 3X80 GR EKO BUT.DILIMLI"16"	2	140,1	97,63	195,26	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
7	1508148	300977	2022-06-10 00:00:00	29478	Serhat GÜMÜŞYAY	11225	DARD TON TON 160 GR "24"	1	13,15	13,36	13,36	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
8	1558829	311040	2023-07-03 00:00:00	12822	Semih Kuzey CALARGÜN	26293	DARDANEL YAGDA ACI 80X3	7	75	31,82	222,74	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
9	1939423	387185	2021-02-07 00:00:00	13357	Erol CANDI	12394	DARD.TON 2X160 GR HAY.BAL.YI "12"	3	113,35	58,36	175,08	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL
10	2076467	414615	2021-11-20 00:00:00	13528	Batuhan Ersin BİCER	19766	DARD TON LIGHT 160X2 GR "12"	7	42,9	30,45	213,15	BALIK	BEYAZ ETLER	TON BALIGI	KONSERVE	DARDANEL

The Google Image search returns a picture of a tuna can.

So, we know we are correct regarding the translation of 'Balik' to 'Fish'.

This is done for all the category values, and changes to the table is made by:

```
UPDATE DATA
SET
    CATEGORY1 = 'FISH'
WHERE
    CATEGORY1 = 'BALIK';
```

Running the previous query of `WHERE CATEGORY1='BALIK'`, yields no rows. Double checking by running

```
SELECT TOP 10 *
FROM DATA
WHERE CATEGORY1='BALIK'
```

Gives:

	ID	ORDERID	DATE_	USERID	NAMESURNAME	ITEMID	ITEMNAME	AMOUNT	UNITPRICE	PRICE	TOTALPRICE	CATEGORY1	CATI
1	1154746	230302	2021-11-10 00:00:00	49195	Şahin EYİLER	19766	DARD TON LIGHT 160X2 GR *12*	1	42,9	30,17	30,17	FISH	BEY
2	1230344	245372	2023-01-01 00:00:00	75608	Hamza Duran FİDANCAN	19766	DARD TON LIGHT 160X2 GR *12*	8	42,9	44,37	354,96	FISH	BEY
3	1466492	292664	2021-01-21 00:00:00	15808	Emir ŞILGIN	25539	DARD TON 185 X 2 GR FASULYELI *12*	2	39,66	19,7	39,4	FISH	BEY
4	1361920	271641	2021-04-24 00:00:00	74853	Muhammet Yiğit GÜREŞÇİ	12393	DARD.TON 3X80 GR HAY.BAL.YI. *16*	6	89,4	51,71	310,26	FISH	BEY
5	1573461	313993	2022-03-20 00:00:00	54786	Erdoğan Engin KUZUBAŞ	19764	DARD TON A.YAGLI 160 GR *24*	4	23,64	21,64	86,56	FISH	BEY
6	1469053	293174	2021-10-09 00:00:00	3941	Hamdi ZABUN	2760	DARD.TON 3X80 GR EKO BUT.DILIMLI*16*	2	140,1	97,63	195,26	FISH	BEY
7	1508148	300977	2022-06-10 00:00:00	29478	Serhat GÜMÜŞYAY	11225	DARD TON TON 160 GR *24*	1	13,15	13,36	13,36	FISH	BEY
8	1558829	311040	2023-07-03 00:00:00	12822	Semih Kuzey CALARGÜN	26293	DARDANEL YAGDA ACI 80X3	7	75	31,82	222,74	FISH	BEY
9	1939423	387185	2021-02-07 00:00:00	13357	Erol CANDI	12394	DARD.TON 2X160 GR HAY.BAL.YI *12*	3	113,35	58,36	175,08	FISH	BEY
10	2076467	414615	2021-11-20 00:00:00	13528	Batuhan Ersin BICER	19766	DARD TON LIGHT 160X2 GR *12*	7	42,9	30,45	213,15	FISH	BEY

So, we are doing good.

Now, I do this for all the category values in the CATEGORY1 column.

Checking the categories again after the update of the table have been done:

```
SELECT DISTINCT CATEGORY1
FROM DATA
ORDER BY CATEGORY1 ASC
```

Gives:

	CATEGORY1
1	BABY
2	BREAKFAST
3	CIGARETTES
4	CLEANING SUPPLIES
5	COLD DRINKS
6	CONFECTIONERY
7	CONSUMPTION
8	COSMETIC
9	DAIRY
10	DETERGENT
11	FABRICS
12	FISH
13	FOOD
14	FRUITS
15	GREENS
16	HOT DRINKS
17	HOUSE
18	MEAT
19	MISCELLANEOUS
20	PAPER
21	TEA-COFFEE-SUGAR
22	TILES
23	TOYS
24	VEGETABLES

Looks good!

Within the category value of ‘CONSUMPTION’, we have ITEMNAMES suggesting products like rice, sugar, bulgur wheat, cinnamon, and chickpeas.

Looking at samples of ITEMNAMES within each category, I discover that the value FABRICS (originally KUMES) is indeed ‘POULTRY’. I update the value in the table for all values of ‘FABRICS’ to ‘POULTRY’.

Within the category value of ‘FOOD’, we find items such as ketchup, mayonnaise, chicken nuggets, and hot pepper paste.

The category value ‘GREENS’ contain items such as green onions, white radish, lettuce, etc.

The category value ‘HOUSE’ contains items such as rubber/eraser, floppy disks, backpacks, pens, notebooks, etc.

The category value of ‘HOUSE’ (originally ‘EV’) is renamed ‘STATIONERY’.

The category value of ‘MEAT’ contains semi-finished products such as meatballs, salami, sausage (even from poultry)

From the category value of ‘MISCELLENOUS’ we find items like plastic bags, tiles, bowls, white cheese, stone curbs, toothbrushes, etc.

Under the category value of ‘PAPER’, we find items such as baby wipes, toilet paper, etc.

The category value of ‘TEA-COFFEE-SUGAR’ contains items such as tea, coffee, and sugar in bulk (powder) in contrast to the category value of ‘HOT DRINKS’ where we also can find coffee drinks, but in smaller containers or bags.

The category value of ‘TILES’ contain items such as tiles, outdoor furniture, outdoor grills, waste baskets/trash cans, etc.

The category value is changed to ‘OUTDOOR FURNITURE’ to better reflect the category value.

So, all the category values of *CATEGORY1’ have now been checked and altered.

This are the final values in the category:

	CATEGORY1
1	BABY
2	BREAKFAST
3	CIGARETTES
4	CLEANING SUPPLIES
5	COLD DRINKS
6	CONFECTIONERY
7	CONSUMPTION
8	COSMETIC
9	DAIRY
10	DETERGENT
11	FABRICS
12	FISH
13	FOOD
14	FRUITS
15	GREENS
16	HOT DRINKS
17	MEAT
18	MISCELLANEOUS
19	OUTDOOR FUNRITURE
20	PAPER
21	STATIONERY
22	TEA-COFFEE-SUGAR
23	TOYS
24	VEGETABLES

Having finished translating the top tier category values, CATEGORY1, we now do the same for the under category, i.e. CATEGORY2, of each category value in CATEGORY1.

This is the code we use:

```
SELECT DISTINCT CATEGORY2
FROM DATA
WHERE CATEGORY1='BABY'
ORDER BY CATEGORY2 ASC
```

Gives:

	CATEGORY2
1	BEBE MALZEMELERI
2	BEBEK MALZEMELERI
3	GIYIM
4	HAZIR YEMEK-MAMA
5	IC GIYIM
6	SAGLIK URUNLERI
7	TEKSTIL-GIYIM-AKSESUAR
8	VUCUT-EL BAKIM

These are presumably under categories to the CATEGORY1='BABY'.

I perform the same translation of the table values like I previously did.

CATEGORY1='BABY'

The category values translate to the following:

BEBE MALZEMELERI	BABY SUPPLIES
BEBEK MALZEMELERI	BABY SUPPLIES
GIYIM	BABY CLOTHES
HAZIR YEMEK-MAMA	BABY MEALS
IC GIYIM	BABY UNDERWEAR
SAGLIK URUNLERI	BABY HEALTH PRODUCTS
TEKSTIL-GIYIM-AKSESUAR	BABY TEXTILES-CLOTHING-ACCESSORIES
VUCUT-EL BAKIM	BABY SOAPS

Searching for the difference between 'bebe' and 'bebek' I find that 'bebe' is used in some rural areas, such as Anatolia, instead of the correct formal name for baby, 'bebek'.

Looking at the items, I find very little difference in items (and image searching them on Google) running the SQL query of:

```
SELECT TOP 5000 *
FROM DATA
WHERE (CATEGORY1='BABY' AND CATEGORY2='BEBEK MALZEMELERI')
```

, and

```
SELECT TOP 5000 *
FROM DATA
WHERE (CATEGORY1='BABY' AND CATEGORY2='BEBE MALZEMELERI')
```

Thus, I choose to combine the two under category values into one, ‘BABY SUPPLIES’.

I do this by running the following SQL query:

```
UPDATE DATA
SET
    CATEGORY2 = 'BABY SUPPLIES'
WHERE
    (CATEGORY2 = 'BEBE MALZEMELERI' OR CATEGORY2 = 'BEBEK MALZEMELERI')
```

Here’s the final version of the category values of CATEGORY2 for CATEGORY1=’BABY’:

	CATEGORY2
1	BABY CLOTHES
2	BABY HEALTH PRODUCTS
3	BABY JELLY
4	BABY MEALS
5	BABY SUPPLIES
6	BABY TEXTILES-CLOTHING-ACCESSORIES
7	BABY UNDERWEAR

Here’s the final version of the category values of CATEGORY2 for CATEGORY1=’BREAKFAST’:

	CATEGORY2
1	BAKERY PRODUCTS
2	BUTTER-MARGARINE
3	CEREALS
4	CHOCOLATE
5	EGG-OLIVE
6	HONEY-JAM
7	MILK-YOGURT-CHEESE
8	TRADITIONAL
9	WHIPPED CREAM-SPREADS

For the values of CATEGORY1=’CLEANING SUPPLIES’, we have:

BULASIK YIKAMA	WASHING UP
EV GERECLERI	HOUSEHOLD APPLIANCES
EV TEMIZLEYICILER	HOUSE CLEANERS
EV TEMIZLIK	HOUSE CLEANING
EV TEMIZLIK GERECLERI	HOUSE CLEANING MATERIALS
KOKULAR	SMELLS
TEMIZLIK GERECLERI	CLEANING REQUIREMENTS

Here, we can see that row 2, 3, 4, 5, and 7 are similar. After having checked their items, I decide to make them into a single category value called ‘CLEANING MATERIALS’.

I used the SQL query:

```
UPDATE DATA
SET
    CATEGORY2 = 'CLEANING MATERIALS'
WHERE
    (CATEGORY2 = 'EV GERECLERI' OR CATEGORY2 = 'EV TEMIZLEYICILER' OR CATEGORY2 = 'EV
TEMIZLIK' OR CATEGORY2 = 'TEMIZLIK GERECLERI' OR CATEGORY2 = 'EV TEMIZLIK GERECLERI')
```

I end up with only three under category values for CATEGORY1=’CLEANING SUPPLIES’:

	CATEGORY2
1	CLEANING MATERIALS
2	REFRESHERS
3	SCRUBBING

A reflection on how to proceed

A lot of work translating, makes me want to look at the number of category values we have:

```
SELECT
       COUNT(DISTINCT CATEGORY1)
       ,COUNT(DISTINCT CATEGORY2)
       ,COUNT(DISTINCT CATEGORY3)
       ,COUNT(DISTINCT CATEGORY4)
FROM DATA
```

Gives:

	(No column name)	(No column name)	(No column name)	(No column name)
1	24	68	162	740

So, there are a lot of work that needs to be put in.

Maybe it would be better to take a view of the top values and translate them once visualization is done, but then again if we are going full dashboard, everything needs to be in excellent order.

I have obviously been thinking about an easier, more computational way, of making translations. That can probably be found, but I still would loose making sure everything is ok with the translation, using a human opinion.

After having made translations from Turkish to English for CATEGORY1, CATEGORY2, and CATEGORY3, I have decided to not make changes to CATEGORY4. Having done 25% of all translations, it takes too long time and I believe that the loss of not having CATEGORY4 is not that important to be able to drill down to that granularity.

So, translations are considered being done. I'll keep the column CATEGORY4, but won't use it in a dashboard in data visualization.

6. Verify that the REGIONSs are indeed Turkish country regions

I run the following script to examine unique REGIONS:

```
SELECT DISTINCT REGION
FROM SALES10M.dbo.DATA
```

Gives:

- İç Anadolu
- Marmara
- Güneydoğu Anadolu
- Karadeniz
- Doğu Anadolu
- Akdeniz
- Ege

I'll make a translation for the regions:

İç Anadolu	Central Anatolia Region
Marmara	Marmara Region
Güneydoğu Anadolu	Southeastern Anatolia Region
Karadeniz	Black Sea Region
Doğu Anadolu	Eastern Anatolia Region
Akdeniz	Mediterranean Region
Ege	Aegean Region

It is done by running the SQL script of:

```
UPDATE DATA
SET REGION='Aegean Region'
WHERE REGION='Ege'
```

7. Verify that the CITY are indeed Turkish cities

With the code:

```
SELECT DISTINCT CITY
FROM DATA
```

We can see that we get a selection of 81 cities in Türkiye.

They are all indeed cities in Türkiye.

8. Verify that the towns are indeed smaller settlements than cities

I ran the following code to examine various towns and map them on Google Maps to make sure they are indeed towns in Türkiye.

```
SELECT TOP 5000 *
FROM DATA
```

9. Verify that DISTRICTS are districts within a city or town

Doing the same with DISTRICS, I noticed that they are smaller villages outside or part of a larger city.

10. What the ADDRESSTEXT represents, Point of Sale or customer address?

The addresses seem to be randomly generated addresses Türkiye, but connected to the DISTRICT, TOWN, CITY, REGION.

11. What is the date range in the column 'DATE_'?

Running:

```
SELECT MIN(DATE_), MAX(DATE_)
FROM DATA
```

Gives:

2021-01-01 00:00:00 2023-08-14 00:00:00

So, from 2021 to mid-2023. About 32 months.

Should be enough to get some seasonal data!

12. Change Turkish indicators for male and female to ‘MALE’ and ‘FEMALE’.

Man or Male is Erkek, while woman or female is Kadın.

In the column ‘USERGENDER’ we have either K or E, according to:

```
SELECT DISTINCT USERGENDER
FROM DATA
```

So, we want to update the table with ‘Male’ and ‘Female’ instead.

Since the datatype(size) is VARCHAR(1) for ‘E’ or ‘K’, we want to change the size of the column to 6. We do this by the following SQL script:

```
ALTER TABLE DATA
ALTER COLUMN USERGENDER
varchar(6);
```

Then we update the table by using the following script:

```
UPDATE DATA
SET USERGENDER='Male'
WHERE USERGENDER='E'

UPDATE DATA
SET USERGENDER='Female'
WHERE USERGENDER='K'
```

STEP 4 – SETTING UP BUSINESS QUESTIONS

If I entertain the idea that an investment company would like to invest in a Turkish food retail group, they will ask some questions about sales, which this dataset can help me to answer.

Or maybe it is the Turkish food retail group itself that would like to get some questions answered from their own data.

1. **Sales Performance**:
    - What is the overall trend in sales revenue over time?
    - Which products or product categories generate the highest sales revenue?
    - How do sales vary by location (e.g., region, store)?
    - Are there any seasonal patterns or trends in sales?
2. **Customer Insights**:
    - Who are the top customers in terms of purchase frequency or total spending?
    - What is the average order size, and how has it changed over time?
3. **Product Analysis**:
    - Which products have the highest and lowest sales volume?
    - What is the product turnover rate or inventory turnover rate based on sales data?
4. **Forecasting**:
    - Can we use historical sales data to forecast future sales and demand?
    - What factors (e.g., economic indicators, seasonality) should be considered in forecasting based on sales data?
5. **Customer Retention**:
    - What is the customer retention rate based on repeat purchases and sales data alone?
    - Are there strategies to encourage repeat purchases based on sales history?
6. **Geographic Expansion**:
    - Are there regions or locations where sales potential is not fully realized based on sales data?
    - What is the market penetration in different geographic areas based on sales performance?