SQL for Exploratory Data Analysis

Rochelle Smits-Seemann
UTOUG
March, 2019

1

About Me

MS in experimental psychology

Nonprofit management

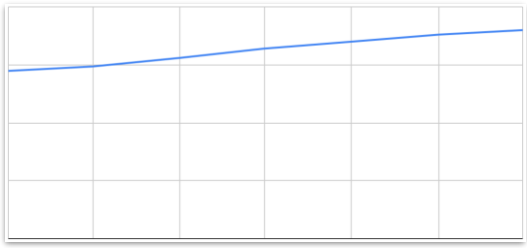
Healthcare research

Higher education

2

Why EDA is important

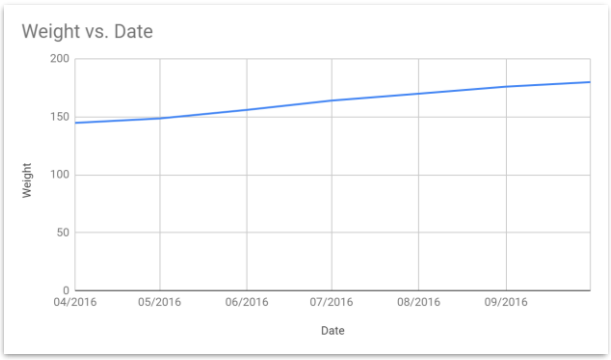
Make a recommendation based on this data:



3

Why EDA is important

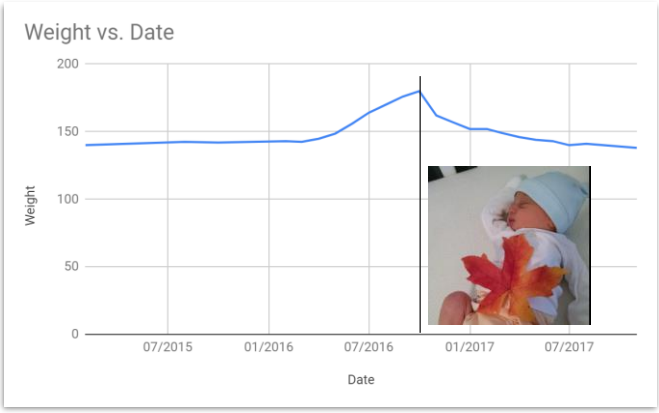
Make a recommendation based on this data:



4

Why EDA is important

Make a recommendation based on this data:



5

Why EDA is important

If you are going to make a visualization

recommendation

predictive model you need to be confident in your data!

6

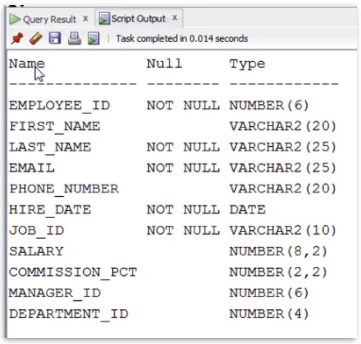
Structurally know your data

Number of columns

Data types

Missing values (Nullable)

```
DESCRIBE TABLE_NAME;
```



Query Result | Script Output

Task completed in 0.014 seconds

Name	Null	Type
EMPLOYEE_ID	NOT NULL	NUMBER(6)
FIRST_NAME		VARCHAR2(20)
LAST_NAME	NOT NULL	VARCHAR2(25)
EMAIL	NOT NULL	VARCHAR2(25)
PHONE_NUMBER		VARCHAR2(20)
HIRE_DATE	NOT NULL	DATE
JOB_ID	NOT NULL	VARCHAR2(10)
SALARY		NUMBER(8,2)
COMMISSION_PCT		NUMBER(2,2)
MANAGER_ID		NUMBER(6)
DEPARTMENT_ID		NUMBER(4)

7

Structurally know your data

```
SELECT *
FROM EMPLOYEES
WHERE ROWNUM < 50
ORDER BY DBMS_RANDOM.VALUE;
```

	EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID	SALARY	COMMISSION_PCT	MANAGER_ID	DEPARTMENT_ID
1	334	Lara	Craft	Lara	(null)	01-MAR-19	IT_PROG	(null)	(null)	(null)	(null)
2	100	Steven	King	SKING	515.123.4567	17-JUN-03	AD_PRES	24000	(null)	(null)	90
3	101	Neena	Kochhar	NKOCHHAR	515.123.4568	21-SEP-05	AD_VP	17000	(null)	100	90
4	102	Lex	De Haan	LDEHAAN	515.123.4569	13-JAN-01	AD_VP	17000	(null)	100	90
5	103	Alexander	Hunold	AHUNOLD	590.423.4567	03-JAN-06	IT_PROG	9000	(null)	102	60
6	104	Bruce	Ernst	BERNST	590.423.4568	21-MAY-07	IT_PROG	6000	(null)	103	60
7	105	David	Austin	DAUSTIN	590.423.4569	25-JUN-05	IT_PROG	4800	(null)	103	60
8	106	Valli	Pataballa	VPATABAL	590.423.4560	05-FEB-06	IT_PROG	4800	(null)	103	60

8

Variable Distributions

```
SELECT MAX(HIRE_DATE)
          ,MIN(HIRE_DATE)
          ,MAX(SALARY)
          ,MIN(SALARY)
          ,MAX(COMMISSION_PCT)
          ,MIN(COMMISSION_PCT)
FROM EMPLOYEES
;
```

	MAX(HIRE_DATE)	MIN(HIRE_DATE)	MAX(SALARY)	MIN(SALARY)	MAX(COMMISSION_PCT)	MIN(COMMISSION_PCT)
1	01-MAR-19	13-JAN-01	24000	2100	0.4	0.1

9

Dates

Precision of date/time variables

```
SELECT ORIENTATION_DATE
FROM ORIENTATIONS
WHERE ROWNUM < 50;
```

	ORIENTATION_DATE
1	21-JUL-17
2	31-JAN-17
3	08-AUG-17
4	12-AUG-15
5	23-AUG-17
6	13-SEP-16
7	13-DEC-16
8	18-JUL-17
9	27-JUN-15
10	28-APR-17
11	11-NOV-14

10

Dates

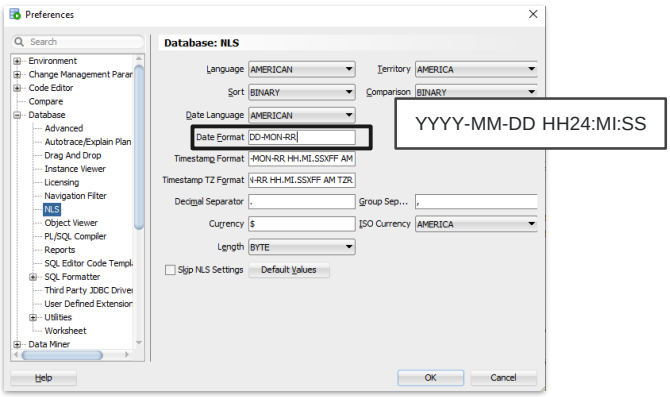
Precision of date/time variables

```
SELECT ORIENTATION_DATE
    ,TO_CHAR(ORIENTATION_DATE, 'DD MON YYYY')
    ,TO_CHAR(ORIENTATION_DATE, 'fmMonth DD, "the year" YYYY')
    ,TO_CHAR(ORIENTATION_DATE, 'fmMonth DD, "the year" YYYY hh:mi:ss AM')
FROM ORIENTATIONS;
```

	ORIENTATION_DATE	TO_CHAR(ORIENTATION_DATE,'DDMONYYYY')	TO_CHAR(ORIENTATION_DATE,'FMMONTHDD,"THEYEAR"YYYY')	TO_CHAR(ORIENTATION_DATE,'FMMONTHDD,"THEYEAR"YYYYHH:MI:SSAM')
1	21-JUL-17	21 JUL 2017	July 21, the year 2017	July 21, the year 2017 12:26:22 PM
2	31-JAN-17	31 JAN 2017	January 31, the year 2017	January 31, the year 2017 12:48:13 PM
3	08-AUG-17	08 AUG 2017	August 8, the year 2017	August 8, the year 2017 4:55:16 PM
4	12-AUG-15	12 AUG 2015	August 12, the year 2015	August 12, the year 2015 2:44:16 PM
5	23-AUG-17	23 AUG 2017	August 23, the year 2017	August 23, the year 2017 10:34:37 PM
6	13-SEP-16	13 SEP 2016	September 13, the year 2016	September 13, the year 2016 5:3:8 PM
7	13-DEC-16	13 DEC 2016	December 13, the year 2016	December 13, the year 2016 10:25:52 AM
8	18-JUL-17	18 JUL 2017	July 18, the year 2017	July 18, the year 2017 12:41:40 PM
9	27-JUN-15	27 JUN 2015	June 27, the year 2015	June 27, the year 2015 1:10:26 PM
10	28-APR-17	28 APR 2017	April 28, the year 2017	April 28, the year 2017 2:32:0 PM
11	11-NOV-14	11 NOV 2014	November 11, the year 2014	November 11, the year 2014 9:4:41 PM

11

Recommendation



12

Dates

Precision of date/time variables

```
SELECT APPLICATION_DATE
    ,TO_CHAR(APPLICATION_DATE, 'fmMonth DD, YYYY hh:mi:ss AM')
FROM APPLICATIONS;
```

	APPLICATION_DATE	TO_CHAR(APPLICATION_DATE,'FMMONTHDD,YYYYHH:MI:SSAM')
1	2003-03-05 22:05:46	March 5, 2003 10:5:46 PM
2	2003-03-05 22:05:46	March 5, 2003 10:5:46 PM
3	2002-10-01 00:00:00	October 1, 2002 12:0:0 AM
4	2003-08-18 00:00:00	August 18, 2003 12:0:0 AM
5	2003-03-05 22:05:46	March 5, 2003 10:5:46 PM
6	2003-03-05 22:05:46	March 5, 2003 10:5:46 PM
7	2016-03-30 00:00:00	March 30, 2016 12:0:0 AM
8	2016-04-11 00:00:00	April 11, 2016 12:0:0 AM

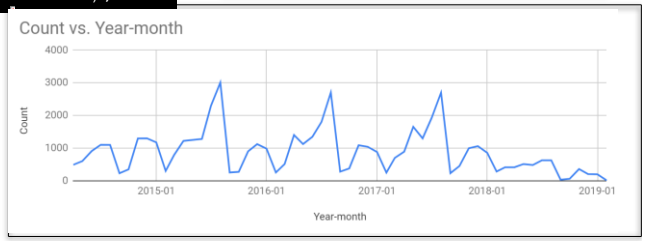
13

Dates

Number of events over time

```
SELECT TO_CHAR(ORIENTATION_DATE, 'YYYY-MM')
       ,COUNT(*)
FROM ORIENTATIONS
GROUP BY TO_CHAR(ORIENTATION_DATE, 'YYYY-MM')
ORDER BY TO_CHAR(ORIENTATION_DATE, 'YYYY-MM');
```

TO_CHAR(ORIENTATION_DATE,'YYYY-MM')	COUNT(*)
2018-04	415
2018-05	514
2018-06	483
2018-07	629
2018-08	629
2018-09	34
2018-10	63
2018-11	361
2018-12	208
2019-01	199
2019-02	11



14

Dates

Number of events over time

```
SELECT TO_CHAR(APPT_DATE, 'YYYY-MM')
          ,DATA_SOURCE
      ,COUNT(*)
FROM APPOINTMENTS
GROUP BY TO_CHAR(APPT_DATE, 'YYYY-MM')
      ,DATA_SOURCE
ORDER BY TO_CHAR(APPT_DATE, 'YYYY-MM');
```

15

Times

What are your business hours?

```
SELECT TO_CHAR(APPOINTMENT_START, 'HH24')
      ,COUNT(*)
FROM APPOINTMENTS
GROUP BY TO_CHAR(APPOINTMENT_START, 'HH24')
ORDER BY TO_CHAR(APPOINTMENT_START, 'HH24');
```

TO_CHAR(APPOINTMENT_START_DATE,'HH24')	COUNT
00	12
01	42
02	35
03	23
04	27
05	25
06	15
07	84
08	2143
09	4120
10	6361
11	8429
12	7975
13	8931
14	8673
15	8636
16	6441
17	4511
18	2230
19	943
20	314
21	21
22	4
23	4

16

Times

Time of day

```
SELECT TO_CHAR(APPOINTMENT_START_DATE, 'HH24') AS APPT_HOUR
      ,TO_CHAR(APPOINTMENT_START_DATE, 'MON') AS APPT_MONTH
      ,COUNT(*) AS COUNT
FROM wsrpmgr.fact_Advising_appointments
GROUP BY TO_CHAR(APPOINTMENT_START_DATE, 'HH24')
        ,TO_CHAR(APPOINTMENT_START_DATE, 'MON')
ORDER BY APPT_MONTH
        ,APPT_HOUR;
```

APPT_HOUR	APPT_MONTH	COUNT
01	APR	2
02	APR	1
03	APR	3
04	APR	5
05	APR	3
06	APR	1
07	APR	3
08	APR	177
09	APR	308
10	APR	501
11	APR	699
12	APR	647
13	APR	777
14	APR	666
15	APR	746
16	APR	534

17

Times

Pivot for consumability

```
SELECT *
FROM (SELECT TO_CHAR(APPOINTMENT_START_DATE, 'HH24') AS APPT_HOUR
          ,TO_CHAR(APPOINTMENT_START_DATE, 'MON') AS APPT_MONTH
          ,COUNT(*) AS COUNT
      FROM wsrpmgr.fact_Advising_appointments
      GROUP BY TO_CHAR(APPOINTMENT_START_DATE, 'HH24')
              ,TO_CHAR(APPOINTMENT_START_DATE, 'MON'))
PIVOT(
  SUM(COUNT)
  FOR APPT_MONTH IN ('JAN', 'FEB', 'MAR', 'APR', 'MAY', 'JUN', 'JUL',
'AUG', 'SEP', 'OCT', 'NOV', 'DEC'))
ORDER BY APPT_HOUR;
```

18

Times

APPT_HOUR	'JAN'	'FEB'	'MAR'	'APR'	'MAY'	'JUN'	'JUL'	'AUG'	'SEP'	'OCT'	'NOV'	'DEC'
00	1	2	1				1	2	1		2	2
01	3	3	7	2	2	3	3	7	3	4	1	4
02	7	2	4	1	4		2	4		4	3	4
03		2	3	3	4	2		5	1	2		1
04	4	3	2	5		2	1	6	1	1	1	1
05	4	1		3	1	3	3	4	2	2		2
06	3	1		1	2	1	2	4		1		
07	25	3	2	3	10	4	4	20		2	7	4
08	355	99	155	177	197	111	182	345	98	95	166	163
09	659	238	252	308	328	258	316	628	228	234	397	274
10	995	353	430	501	551	389	497	984	322	331	597	411
11	1144	480	547	699	699	531	678	1350	443	504	807	547
12	1100	468	553	647	714	520	636	1261	394	437	724	521
13	1169	439	587	777	786	604	745	1395	478	554	830	567
14	1183	452	584	666	714	594	810	1394	426	490	791	569
15	1145	432	570	746	768	586	770	1407	398	482	737	595
16	907	322	373	534	573	511	576	1131	300	335	465	414
17	624	222	274	357	434	356	441	749	203	197	361	293
18	264	95	128	192	221	184	204	411	112	104	178	137
19	111	26	52	93	90	88	110	164	30	56	75	48
20	27	8	15	25	25	19	48	71	11	11	31	23
21				5	1	3	3	2	3	2	1	1
22				1					1			2
23		1		1	1							1

19

Dates

Duration of events

```
SELECT APPOINTMENT_START,
    APPOINTMENT_END,
    APPOINTMENT_END -  APPOINTMENT_START AS DIFF
FROM APPOINTMENTS;
```

APPOINTMENT_START	APPOINTMENT_END	DIFF
11/16/2018 12:04:00 PM	11/16/2018 12:15:00 PM	0.00763888888888889
4/6/2018 10:31:00 AM	4/6/2018 10:45:00 AM	0.00972222222222222
3/19/2018 10:28:00 AM	3/19/2018 11:00:00 AM	0.0222222222222222
6/19/2018 12:58:00 PM	6/19/2018 1:05:00 PM	0.00486111111111111
8/13/2018 1:15:00 PM	8/13/2018 1:32:00 PM	0.0118055555555556
12/4/2018 1:30:00 PM	12/4/2018 2:20:00 PM	0.0347222222222222
7/25/2018 11:15:00 AM	7/25/2018 11:55:00 AM	0.0277777777777778

20

Dates

Duration of events

```
SELECT APPOINTMENT_START,
     APPOINTMENT_END,
     APPOINTMENT_END -  APPOINTMENT_START AS DIFF,
     (APPOINTMENT_END -  APPOINTMENT_START)*24*60 AS LENGTH_MIN
FROM APPOINTMENTS;
```

APPOINTMENT_START	APPOINTMENT_END	DIFF	LENGTH_MINUTES
11/16/2018 12:04:00 PM	11/16/2018 12:15:00 PM	0.00763888888888889	11
4/6/2018 10:31:00 AM	4/6/2018 10:45:00 AM	0.00972222222222222	14
3/19/2018 10:28:00 AM	3/19/2018 11:00:00 AM	0.0222222222222222	32
6/19/2018 12:58:00 PM	6/19/2018 1:05:00 PM	0.00486111111111111	7
8/13/2018 1:15:00 PM	8/13/2018 1:32:00 PM	0.0118055555555556	17
12/4/2018 1:30:00 PM	12/4/2018 2:20:00 PM	0.0347222222222222	50

21

Dates

Assumptions to check

- Do all appointments start before they end?
- Are any appointments missing start or end datetimes?

```
SELECT *
FROM APPOINTMENTS
WHERE APPOINTMENT_END < APPOINTMENT_START;

SELECT *
FROM APPOINTMENTS
WHERE APPOINTMENT_END IS NULL
OR APPOINTMENT_START IS NULL;
```

22

Dates

Realistic appointment lengths

```
SELECT ROUND((APPOINTMENT_END_DATE -
APPOINTMENT_START_DATE)*24*60, -1) AS LENGTH_MINUTES
     ,COUNT(*)
FROM APPOINTMENTS
GROUP BY ROUND((APPOINTMENT_END_DATE -
APPOINTMENT_START_DATE)*24*60, -1)
ORDER BY COUNT(*) DESC;
```

LENGTH_MINUTES	COUNT(*)
20	14197
10	13348
30	10138
0	7545
120	6606
40	5692
50	3708

23

Dates

Automatic time-outs

```
SELECT (APPOINTMENT_END_DATE -  APPOINTMENT_START_DATE)*24*60 AS
LENGTH_MIN
     ,COUNT(*)
     ,COUNT(*) OVER()
FROM TUTOR_APPOINTMENTS
GROUP BY ROUND((APPOINTMENT_END_DATE -
APPOINTMENT_START_DATE)*24*60, -1)
ORDER BY COUNT(*) DESC;
```

LENGTH_MIN	COUNT(*)	COUNT(*)OVER()
8	997	2933
60	34	2933
0.05	13	2933
0.1	11	2933
0.0166666666666667	11	2933
0.0333333333333333	11	2933
0.0666666666666667	10	2933
11.5666666666667	7	2933
30	7	2933

24

Reasonable values

What is a reasonable range and distribution of values?

```
SELECT ROUND(SYSDATE - BIRTHDATE, -1) AS AGE_TODAY
     ,COUNT(*)
FROM EMPLOYEES
GROUP BY ROUND(SYSDATE - BIRTHDATE, -1)
ORDER BY ROUND(SYSDATE - BIRTHDATE, -1);
```

ROUND(AGE_TODAY,-1)	COUNT(*)
10	8
20	1820
30	6438
40	21667
50	16611
60	12246
70	5308
80	1756
90	680
100	150
110	11
120	30
	3274

25

Weird values

TEST	MIN(TEST_SCORE)	MAX(TEST_SCORE)
SAT Test Date	0000	1811
SAT Total Score	1000	910
SAT Verbal	280	740
SAT Writing	18	790
Qualtrics English Date	1901	1902
Qualtrics English Score	0000	1010
Qualtrics Math Date	1901	1902
Qualtrics Math Score	0950	1060

26

Weird values

Student name - “Ima Student”

Duplicate records

Name	ID
John Smith	2345
John Smith DO NOT USE	1346

27

Summary

Context is king

Structural knowledge is only the beginning

Variable precision is important

Use business knowledge to check ranges and distributions

Check your assumptions

28

SQL for Exploratory Data Analysis

Rochelle Smits-Seemann
UTOUG
March, 2019

29

Reference

Great blog post - https://towardsdatascience.com/data-science-foundations-know-your-data-really-really-know-it-a6bb97eb991c?sk=42f1c02883e744df7dbb618373312244

30