BUAN 6320 -
Database Foundations for Business Analytics

Presented to Professor Farzad Kamalzadeh

New York Citywide Payroll Data (Fiscal Year)

Source: https://usamagazine.net/different-types-of-payroll-management-systems-used-by-organizations/

PRESENTED BY:

Name: Andrew Brocious	**Net ID**: AWB170030
Name: Bo Zhang	**Net ID**: BXZ110730
Name: Shwetal Malviya	**Net ID**: SXM210157
Name: Taufeeq Samnani	**Net ID**: TMS210000
Name: Sookie (Yixiao) Ji	**Net ID**: YXJ210010

TABLE OF CONTENTS

Tables in the DATABASE

Database Schema Design

1. Create Table Statements

BUSINESS UNDERSTANDING

For any business, payroll generation is an important step. The data collected in this process is used to maintain relational database to keep track of employees, their relation to organization, timesheets of work done and payment generation.

This dataset is hosted by the City of New York (NY). The curiosity of NY City's population to understand how the City's budget is being spent on overtime pay and salary for all municipal employees led to the collection of data here. The data was collected in NY City's Personnel Management System ("PMS") and was input by different user agencies. The collected data can be utilized to make analysis about how much of the NY City's budget is allotted to overtime and how the City's financial resources are being used.

Note to the reader of this data: Only the gross and final base salary of an employee at the end of the fiscal year is captured here. Increments of salary increase received over the period of any one fiscal year is not reflected.

DATA UNDERSTANDING

The NY City has an open data platform which can be found here. This information is updated according to the amount of data that is brought in. The update frequency for this dataset is annually.

Each record represents the following statistics for every city employee: First Name, Last Name, Agency, Middle Initial, Agency Start Date, Job Title Description, Work Location Borough, Leave Status as of the close of the FY (June 30th), Base Salary, Pay Basis, Regular Hours, Regular Gross Paid, Overtime Hours, Total Overtime Paid and Total Other Compensation (i.e., lump sum and/or retro payments).

The major data types used in the dataset are as follows:

- String
- Decimal
- Integer
- Other

Column Name	Column Description	Additional Notes
Payroll Description	The payroll agency that the employee works for	
Last Name	Last name of employee	
First Name	First name of employee	
Middle Initial	Middle initial of employee	
Agency Start Date	Date which employee began working for their current agency	
Work Location Borough	Borough of employee's primary work location	
Leave Status as of Jun 30	Status of employee as of the close of the relevant fiscal year: Active, Ceased, or On Leave	
Title Description	Civil service title description of the employee	
Base Salary	Base Salary assigned to the employee	Base Salary represents the amount the job pays (not necessarily what was earned) and not including any other pay (differentials, lump sums, uniform allowance, meal allowance, retroactive pay increases, settlement amounts, etc.) or overtime
Pay Basis	Lists whether the employee is paid on an hourly, per diem or annual basis	
Regular Hours	Number of regular hours employee worked in the fiscal year	This does not include overtime hours
Regular Gross Paid	The amount paid to the employee for base salary during the fiscal year	Regular gross paid represents actual base salary during reporting period, which is the portion of the person's annual salary paid before deductions are calculated\withheld. This does not include overtime pay or other compensation and does not reflect the after-tax amount or net pay. Total gross pay is calculated by adding columns L, N and O.
OT Hours	Overtime Hours worked by employee in the fiscal year	
Total OT Paid	Total overtime pay paid to the employee in the fiscal year	
Total Other Pay	Includes any compensation in addition to gross salary and overtime pay, i.e., Differentials, lump sums, uniform allowance, meal allowance, retroactive pay increases, settlement amounts, and bonus pay, if applicable.	Not every employee will have a value in this field. For those employees with no other pay, earnings will be stated as $0

TABLES IN THE DATABASE

Employments Table

The Employment table is the most important table in the entire database. It is both directly and indirectly related to every other table in the data model. It stores borough_id, person_id, agency_id, and title_id of all the employees, taken from rest of the tables in the database.

Titles Table

Titles table contains description about various job titles of employees.

Agencies Table

Agencies table captures information about different agencies with their names and IDs.

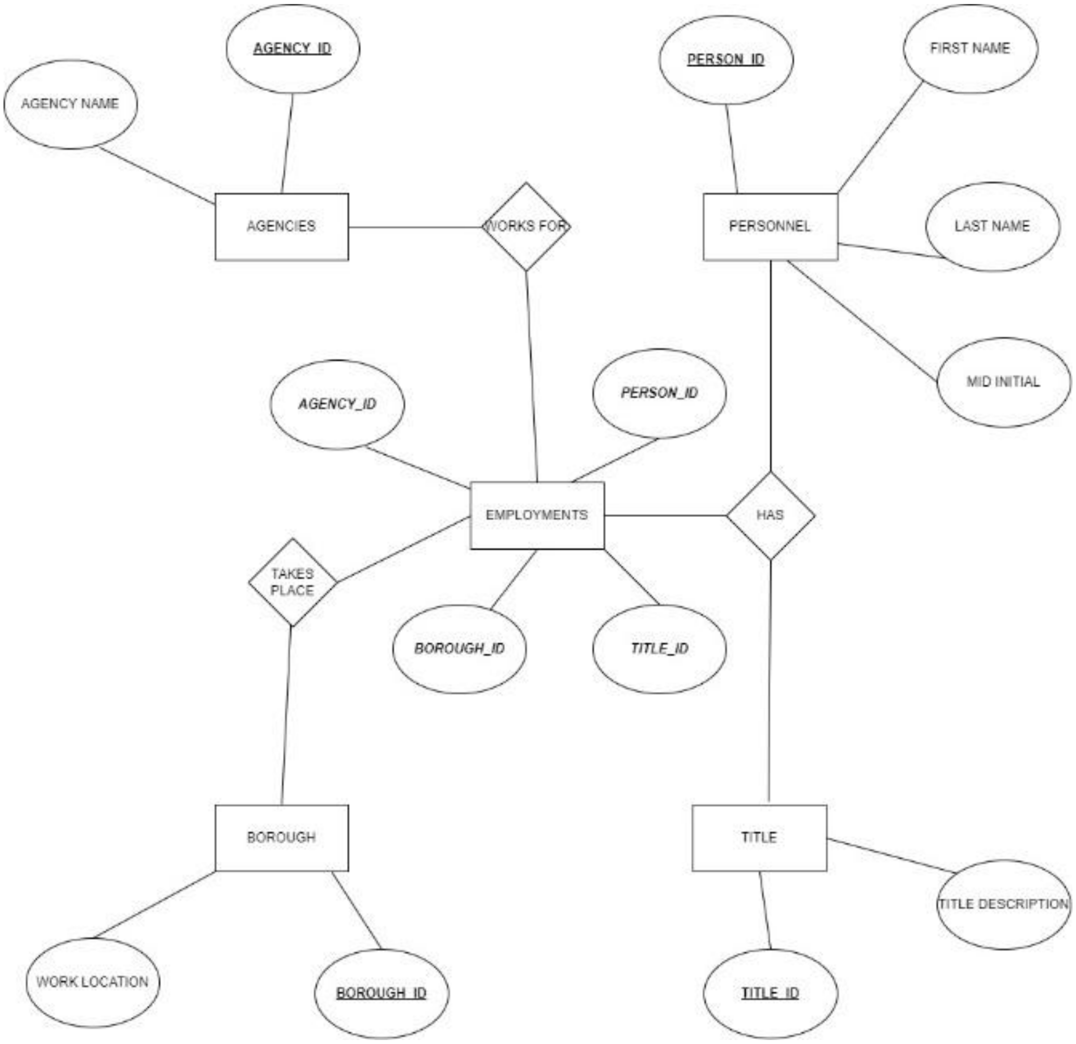
Personnel Table

Personnel table in the database stores the details of employees like First, Last and Middle names.

Boroughs Table

Boroughs table gives us the insights on work location of employees. It is identified by borough_id.

ENTITY RELATIONSHIP DIAGRAM



DATABASE SCHEMA DESIGN

1. CREATE TABLE STATEMENTS

a) Employments Table

```
Create Table Employments
Select a.borough_id, b.person_id, c.agency_id, d.title_id, e.* From
boroughs as a, personnel as b, agencies as c, titles as d, pay as e
Where (a.borough = e.`Work.Location.Borough`)
And (b.first_name = e.`First.Name`)
And (b.last_name = e.`Last.Name`)
And (b.mid_init = e.`Mid.Init`)
And (c.agency = e.`Agency.Name`)
And (d.title = e.`Title.Description`);
```

b) Titles Table

```
Create Table titles
(title_id Int Not Null auto_increment Primary Key)
Select Distinct Replace(Replace(`Title.Description`, '*', ' '), '?', ' ') as title
From pay Order By title;
```

c) Agencies Table

```
Create Table agencies
(agency_id Int Not Null auto_increment Primary Key)
Select Distinct `Agency.Name` as agency
From pay Order By agency;
```

d) Personnel Table

```
Create Table personnel
(person_id Int Not Null auto_increment Primary Key)
Select Distinct `First Name` as first_name, `Mid Init` as mid_init, `Last Name` as last_name
From pay Order By `Last Name`;
```

e) Boroughs Table

```
Create Table boroughs
(borough_id Int Not Null auto_increment Primary Key)
Select Distinct `Work Location Borough` as borough
From Pay Order By borough;
```

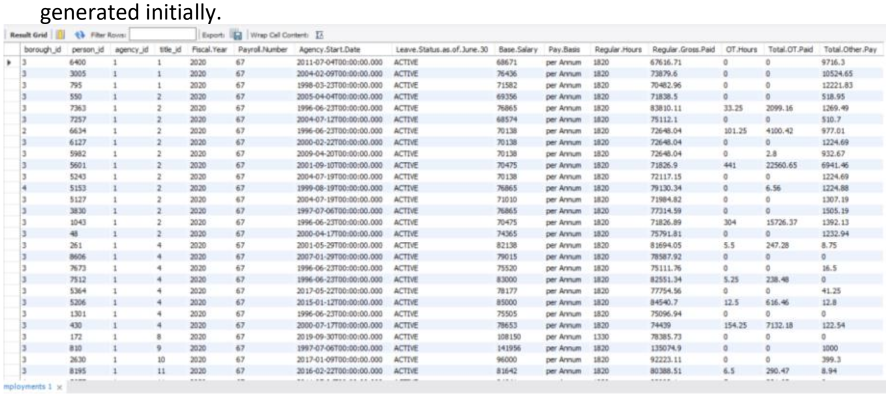
2. ALTER EMPLOYMENT TABLE STATEMENTS

```
Alter Table Employments Add Foreign Key (borough_id) References boroughs(borough_id);
Alter Table Employments Add Foreign Key (person_id) References personnel(person_id);
Alter Table Employments Add Foreign Key (agency_id) References agencies(agency_id);
Alter Table Employments Add Foreign Key (title_id) References titles(title_id);
Alter Table Employments Drop Column `First.Name`;
Alter Table Employments Drop Column `Mid.Init`;
Alter Table Employments Drop Column `Last.Name`;
Alter Table Employments Drop Column `Title.Description`;
Alter Table Employments Drop Column `Agency.Name`;
Alter Table Employments Drop Column `Work.Location.Borough`;
```

DATA CLEANSE

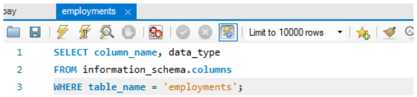
Employments table

1. We have identified the most data errors in Employments Table. This is the screenshot of Employments table when it's generated initially.

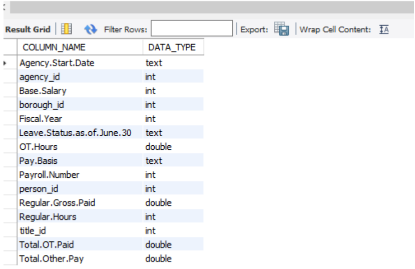


borough_id	person_id	agency_id	title_id	Fiscal.Year	Payroll.Number	Agency.Start.Date	Leave.Status.as.of.June.30	Base.Salary	Pay.Basis	Regular.Hours	Regular.Gross.Paid	OT.Hours	Total.OT.Paid	Total.Other.Pay
3	6400	1	1	2020	67	2011-07-04T00:00:00.000	ACTIVE	68671	per Annum	1820	67616.71	0	0	9716.3
3	3005	1	1	2020	67	2004-02-09T00:00:00.000	ACTIVE	76436	per Annum	1820	73879.6	0	0	10524.65
3	795	1	1	2020	67	1998-03-23T00:00:00.000	ACTIVE	71582	per Annum	1820	70482.96	0	0	12221.83
3	550	1	2	2020	67	2005-04-04T00:00:00.000	ACTIVE	69356	per Annum	1820	71838.5	0	0	518.95
3	7363	1	2	2020	67	1996-06-23T00:00:00.000	ACTIVE	76865	per Annum	1820	83810.11	33.25	2099.16	1269.49
3	7257	1	2	2020	67	2004-07-12T00:00:00.000	ACTIVE	68574	per Annum	1820	75112.1	0	0	510.7
2	6634	1	2	2020	67	1996-06-23T00:00:00.000	ACTIVE	70138	per Annum	1820	72648.04	101.25	4100.42	977.01
3	6127	1	2	2020	67	2000-02-22T00:00:00.000	ACTIVE	70138	per Annum	1820	72648.04	0	0	1224.69
3	5982	1	2	2020	67	2009-04-20T00:00:00.000	ACTIVE	70138	per Annum	1820	72648.04	0	2.8	932.67
3	5601	1	2	2020	67	2001-09-10T00:00:00.000	ACTIVE	70475	per Annum	1820	71826.9	441	22560.65	6941.46
3	5243	1	2	2020	67	2004-07-19T00:00:00.000	ACTIVE	70138	per Annum	1820	72117.15	0	0	1224.69
4	5153	1	2	2020	67	1999-08-19T00:00:00.000	ACTIVE	76865	per Annum	1820	79130.34	0	6.56	1224.88
3	5127	1	2	2020	67	2004-07-19T00:00:00.000	ACTIVE	71010	per Annum	1820	71984.82	0	0	1307.19
3	3830	1	2	2020	67	1997-07-06T00:00:00.000	ACTIVE	76865	per Annum	1820	77314.59	0	0	1505.19
3	1043	1	2	2020	67	1996-06-23T00:00:00.000	ACTIVE	70475	per Annum	1820	71826.89	304	15726.37	1392.13
3	48	1	2	2020	67	2000-04-17T00:00:00.000	ACTIVE	74365	per Annum	1820	75791.81	0	0	1232.94
3	261	1	4	2020	67	2001-05-29T00:00:00.000	ACTIVE	82138	per Annum	1820	81694.05	5.5	247.28	8.75
3	8606	1	4	2020	67	2007-01-29T00:00:00.000	ACTIVE	79015	per Annum	1820	78587.92	0	0	0
3	7673	1	4	2020	67	1996-06-23T00:00:00.000	ACTIVE	75520	per Annum	1820	75111.76	0	0	16.5
3	7512	1	4	2020	67	1996-06-23T00:00:00.000	ACTIVE	83000	per Annum	1820	82551.34	5.25	238.48	0
3	5364	1	4	2020	67	2017-05-22T00:00:00.000	ACTIVE	78177	per Annum	1820	77754.56	0	0	41.25
3	5206	1	4	2020	67	2015-01-12T00:00:00.000	ACTIVE	85000	per Annum	1820	84540.7	12.5	616.46	12.8
3	1301	1	4	2020	67	1996-06-23T00:00:00.000	ACTIVE	75505	per Annum	1820	75096.94	0	0	0
3	430	1	4	2020	67	2000-07-17T00:00:00.000	ACTIVE	78653	per Annum	1820	74439	154.25	7132.18	122.54
3	172	1	8	2020	67	2019-09-30T00:00:00.000	ACTIVE	108150	per Annum	1330	78385.73	0	0	0
3	810	1	9	2020	67	1997-07-06T00:00:00.000	ACTIVE	141956	per Annum	1820	135074.9	0	0	1000
3	2630	1	10	2020	67	2017-01-09T00:00:00.000	ACTIVE	96000	per Annum	1820	92223.11	0	0	399.3
3	8195	1	11	2020	67	2016-02-22T00:00:00.000	ACTIVE	81642	per Annum	1820	80388.51	6.5	290.47	8.94

2. First check the data type of each column:

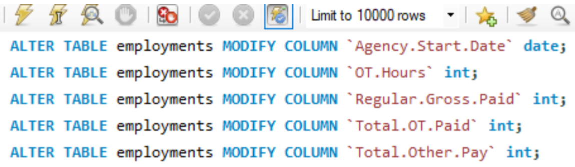


```
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'employments';
```

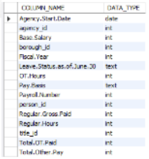


COLUMN_NAME	DATA_TYPE
Agency.Start.Date	text
agency_id	int
Base.Salary	int
borough_id	int
Fiscal.Year	int
Leave.Status.as.of.June.30	text
OT.Hours	double
Pay.Basis	text
Payroll.Number	int
person_id	int
Regular.Gross.Paid	double
Regular.Hours	int
title_id	int
Total.OT.Paid	double
Total.Other.Pay	double

3. Modify column types:



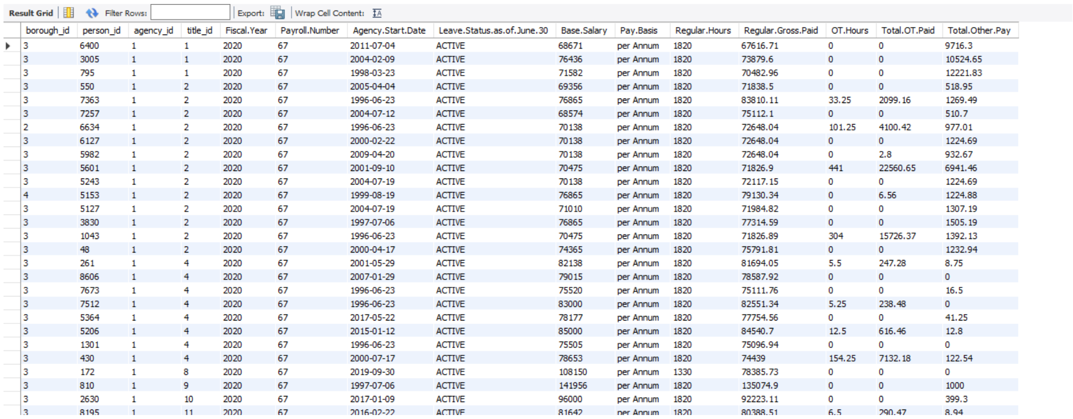
```
ALTER TABLE employments MODIFY COLUMN `Agency.Start.Date` date;
ALTER TABLE employments MODIFY COLUMN `OT.Hours` int;
ALTER TABLE employments MODIFY COLUMN `Regular.Gross.Paid` int;
ALTER TABLE employments MODIFY COLUMN `Total.OT.Paid` int;
ALTER TABLE employments MODIFY COLUMN `Total.Other.Pay` int;
```



COLUMN_NAME	DATA_TYPE
Agency.Start.Date	date
agency_id	int
Base.Salary	int
borough_id	int
Fiscal.Year	int
Leave.Status.as.of.June.30	text
OT.Hours	int
Pay.Basis	text
Payroll.Number	int
person_id	int
Regular.Gross.Paid	int
Regular.Hours	int
title_id	int
Total.OT.Paid	int
Total.Other.Pay	int

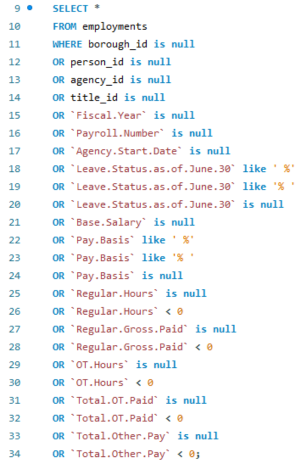
4. Update the table with correct date format for Agency Start Date column using:

```
Update Employments Set `Agency.Start.Date` = Convert(Substring_Index(`Agency.Start.Date`, 'T', 1), date);
```

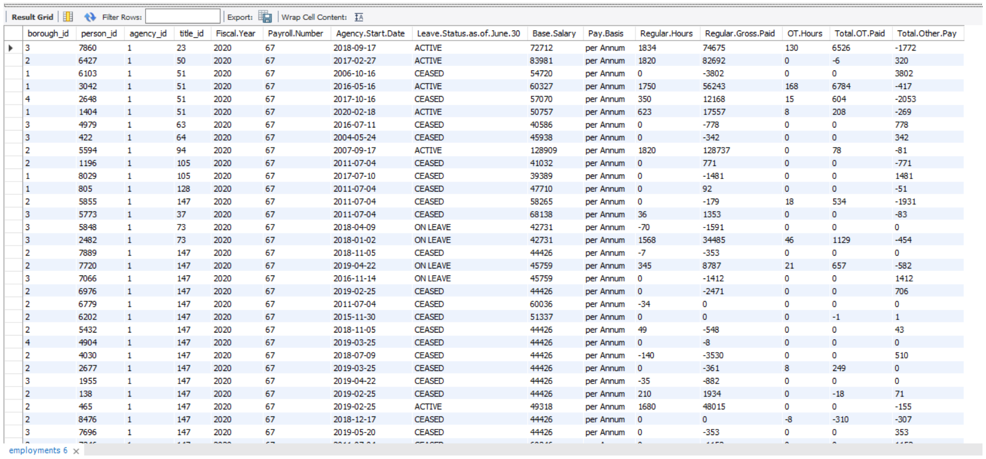


borough_id	person_id	agency_id	title_id	Fiscal.Year	Payroll.Number	Agency.Start.Date	Leave.Status.as.of.June.30	Base.Salary	Pay.Basis	Regular.Hours	Regular.Gross.Paid	OT.Hours	Total.OT.Paid	Total.Other.Pay
3	6400	1	1	2020	67	2011-07-04	ACTIVE	68671	per Annum	1820	67616.71	0	0	9716.3
3	3005	1	1	2020	67	2004-02-09	ACTIVE	76436	per Annum	1820	73879.6	0	0	10524.65
3	795	1	1	2020	67	1998-03-23	ACTIVE	71582	per Annum	1820	70482.96	0	0	12221.83
3	550	1	2	2020	67	2005-04-04	ACTIVE	69356	per Annum	1820	71838.5	0	0	518.95
3	7363	1	2	2020	67	1996-06-23	ACTIVE	76865	per Annum	1820	83810.11	33.25	2099.16	1269.49
3	7257	1	2	2020	67	2004-07-12	ACTIVE	68574	per Annum	1820	75112.1	0	0	510.7
2	6634	1	2	2020	67	1996-06-23	ACTIVE	70138	per Annum	1820	72648.04	101.25	4100.42	977.01
3	6127	1	2	2020	67	2000-02-22	ACTIVE	70138	per Annum	1820	72648.04	0	0	1224.69
3	5982	1	2	2020	67	2009-04-20	ACTIVE	70138	per Annum	1820	72648.04	0	2.8	932.67
3	5601	1	2	2020	67	2001-09-10	ACTIVE	70475	per Annum	1820	71826.9	441	22560.65	6941.46
3	5243	1	2	2020	67	2004-07-19	ACTIVE	70138	per Annum	1820	72117.15	0	0	1224.69
4	5153	1	2	2020	67	1999-08-19	ACTIVE	76865	per Annum	1820	79130.34	0	6.56	1224.88
3	5127	1	2	2020	67	2004-07-19	ACTIVE	71010	per Annum	1820	71984.82	0	0	1307.19
3	3830	1	2	2020	67	1997-07-06	ACTIVE	76865	per Annum	1820	77314.59	0	0	1505.19
3	1043	1	2	2020	67	1996-06-23	ACTIVE	70475	per Annum	1820	71826.89	304	15726.37	1392.13
3	48	1	2	2020	67	2000-04-17	ACTIVE	74365	per Annum	1820	75791.81	0	0	1232.94
3	261	1	4	2020	67	2001-05-29	ACTIVE	82138	per Annum	1820	81694.05	5.5	247.28	8.75
3	8606	1	4	2020	67	2007-01-29	ACTIVE	79015	per Annum	1820	78587.92	0	0	0
3	7673	1	4	2020	67	1996-06-23	ACTIVE	75520	per Annum	1820	75111.76	0	0	16.5
3	7512	1	4	2020	67	1996-06-23	ACTIVE	83000	per Annum	1820	82551.34	5.25	238.48	0
3	5364	1	4	2020	67	2017-05-22	ACTIVE	78177	per Annum	1820	77754.56	0	0	41.25
3	5206	1	4	2020	67	2015-01-12	ACTIVE	85000	per Annum	1820	84540.7	12.5	616.46	12.8
3	1301	1	4	2020	67	1996-06-23	ACTIVE	75505	per Annum	1820	75096.94	0	0	0
3	430	1	4	2020	67	2000-07-17	ACTIVE	78653	per Annum	1820	74439	154.25	7132.18	122.54
3	172	1	8	2020	67	2019-09-30	ACTIVE	108150	per Annum	1330	78385.73	0	0	0
3	810	1	9	2020	67	1997-07-06	ACTIVE	141956	per Annum	1820	135074.9	0	0	1000
3	2630	1	10	2020	67	2017-01-09	ACTIVE	96000	per Annum	1820	92223.11	0	0	399.3
3	8195	1	11	2020	67	2016-02-22	ACTIVE	81642	per Annum	1820	80388.51	6.5	290.47	8.94

5. Sort out the data that are null or invalid in both character columns and numerical columns using the following queries:

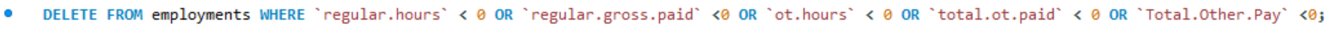


```sql
SELECT *
FROM employments
WHERE borough_id is null
OR person_id is null
OR agency_id is null
OR title_id is null
OR `Fiscal.Year` is null
OR `Payroll.Number` is null
OR `Agency.Start.Date` is null
OR `Leave.Status.as.of.June.30` like ' %'
OR `Leave.Status.as.of.June.30` like '% '
OR `Leave.Status.as.of.June.30` is null
OR `Base.Salary` is null
OR `Pay.Basis` like ' %'
OR `Pay.Basis` like '% '
OR `Pay.Basis` is null
OR `Regular.Hours` is null
OR `Regular.Hours` < 0
OR `Regular.Gross.Paid` is null
OR `Regular.Gross.Paid` < 0
OR `OT.Hours` is null
OR `OT.Hours` < 0
OR `Total.OT.Paid` is null
OR `Total.OT.Paid` < 0
OR `Total.Other.Pay` is null
OR `Total.Other.Pay` < 0;
```



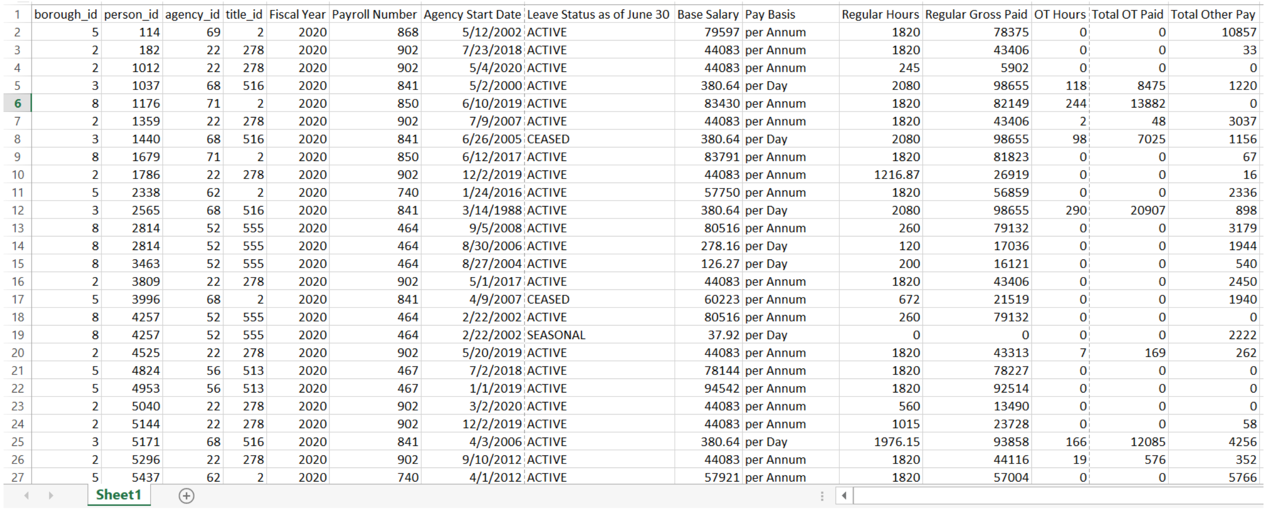
borough_id	person_id	agency_id	title_id	Fiscal.Year	Payroll.Number	Agency.Start.Date	Leave.Status.as.of.June.30	Base.Salary	Pay.Basis	Regular.Hours	Regular.Gross.Paid	OT.Hours	Total.OT.Paid	Total.Other.Pay
3	7860	1	23	2020	67	2018-09-17	ACTIVE	72712	per Annum	1834	74675	130	6526	-1772
2	6427	1	50	2020	67	2017-02-27	ACTIVE	83981	per Annum	1820	82692	0	-6	320
1	6103	1	51	2020	67	2006-10-16	CEASED	54720	per Annum	0	-3802	0	0	3802
1	3042	1	51	2020	67	2016-05-16	ACTIVE	60327	per Annum	1750	56243	168	6784	-417
4	2648	1	51	2020	67	2017-10-16	CEASED	57070	per Annum	350	12168	15	604	-2053
1	1404	1	51	2020	67	2020-02-18	ACTIVE	50757	per Annum	623	17557	8	208	-269
3	4979	1	63	2020	67	2016-07-11	CEASED	40586	per Annum	0	-778	0	0	778
3	422	1	64	2020	67	2004-05-24	CEASED	45938	per Annum	0	-342	0	0	342
2	5594	1	94	2020	67	2007-09-17	ACTIVE	128909	per Annum	1820	128737	0	78	-81
2	1196	1	105	2020	67	2011-07-04	CEASED	41032	per Annum	0	771	0	0	-771
1	8029	1	105	2020	67	2017-07-10	CEASED	39389	per Annum	0	-1481	0	0	1481
1	805	1	128	2020	67	2011-07-04	CEASED	47710	per Annum	0	92	0	0	-51
2	5855	1	147	2020	67	2011-07-04	CEASED	58265	per Annum	0	-179	18	534	-1931
3	5773	1	37	2020	67	2011-07-04	CEASED	68138	per Annum	36	1353	0	0	-83
3	5848	1	73	2020	67	2018-04-09	ON LEAVE	42731	per Annum	-70	-1591	0	0	0
3	2482	1	73	2020	67	2018-01-02	ON LEAVE	42731	per Annum	1568	34485	46	1129	-454
2	7889	1	147	2020	67	2018-11-05	CEASED	44426	per Annum	-7	-353	0	0	0
2	7720	1	147	2020	67	2019-04-22	ON LEAVE	45759	per Annum	345	8787	21	657	-582
3	7066	1	147	2020	67	2016-11-14	ON LEAVE	45759	per Annum	0	-1412	0	0	1412
2	6976	1	147	2020	67	2019-02-25	CEASED	44426	per Annum	0	-2471	0	0	706
2	6779	1	147	2020	67	2011-07-04	CEASED	60036	per Annum	-34	0	0	0	0
2	6202	1	147	2020	67	2015-11-30	CEASED	51337	per Annum	0	0	0	-1	1
2	5432	1	147	2020	67	2018-11-05	CEASED	44426	per Annum	49	-548	0	0	43
4	4904	1	147	2020	67	2019-03-25	CEASED	44426	per Annum	0	-8	0	0	0
2	4030	1	147	2020	67	2018-07-09	CEASED	44426	per Annum	-140	-3530	0	0	510
2	2677	1	147	2020	67	2019-03-25	CEASED	44426	per Annum	0	-361	8	249	0
3	1955	1	147	2020	67	2019-04-22	CEASED	44426	per Annum	-35	-882	0	0	0
2	138	1	147	2020	67	2019-02-25	CEASED	44426	per Annum	210	1934	0	-18	71
2	465	1	147	2020	67	2019-02-25	ACTIVE	49318	per Annum	1680	48015	0	0	-155
2	8476	1	147	2020	67	2018-12-17	CEASED	44426	per Annum	0	0	-8	-310	-307
3	7696	1	147	2020	67	2019-05-20	CEASED	44426	per Annum	0	-353	0	0	353

6. Since there are negative values for Regular Hours, Regular Gross Paid, OT Hours, Total OT Paid and Total Other Pay, we will drop them using:



```sql
DELETE FROM employments WHERE `regular.hours` < 0 OR `regular.gross.paid` <0 OR `ot.hours` < 0 OR `total.ot.paid` < 0 OR `Total.Other.Pay` <0;
```

7. Now this is a screenshot of Employment table exported after data cleanse.

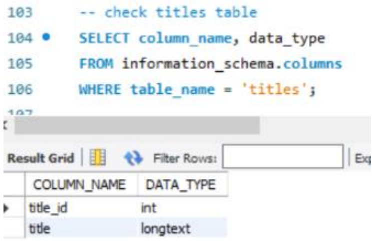


1	borough_id	person_id	agency_id	title_id	Fiscal Year	Payroll Number	Agency Start Date	Leave Status as of June 30	Base Salary	Pay Basis	Regular Hours	Regular Gross Paid	OT Hours	Total OT Paid	Total Other Pay
2	5	114	69	2	2020	868	5/12/2002	ACTIVE	79597	per Annum	1820	78375	0	0	10857
3	2	182	22	278	2020	902	7/23/2018	ACTIVE	44083	per Annum	1820	43406	0	0	33
4	2	1012	22	278	2020	902	5/4/2020	ACTIVE	44083	per Annum	245	5902	0	0	0
5	3	1037	68	516	2020	841	5/2/2000	ACTIVE	380.64	per Day	2080	98655	118	8475	1220
6	8	1176	71	2	2020	850	6/10/2019	ACTIVE	83430	per Annum	1820	82149	244	13882	0
7	2	1359	22	278	2020	902	7/9/2007	ACTIVE	44083	per Annum	1820	43406	2	48	3037
8	3	1440	68	516	2020	841	6/26/2005	CEASED	380.64	per Day	2080	98655	98	7025	1156
9	8	1679	71	2	2020	850	6/12/2017	ACTIVE	83791	per Annum	1820	81823	0	0	67
10	2	1786	22	278	2020	902	12/2/2019	ACTIVE	44083	per Annum	1216.87	26919	0	0	16
11	5	2338	62	2	2020	740	1/24/2016	ACTIVE	57750	per Annum	1820	56859	0	0	2336
12	3	2565	68	516	2020	841	3/14/1988	ACTIVE	380.64	per Day	2080	98655	290	20907	898
13	8	2814	52	555	2020	464	9/5/2008	ACTIVE	80516	per Annum	260	79132	0	0	3179
14	8	2814	52	555	2020	464	8/30/2006	ACTIVE	278.16	per Day	120	17036	0	0	1944
15	8	3463	52	555	2020	464	8/27/2004	ACTIVE	126.27	per Day	200	16121	0	0	540
16	2	3809	22	278	2020	902	5/1/2017	ACTIVE	44083	per Annum	1820	43406	0	0	2450
17	5	3996	68	2	2020	841	4/9/2007	CEASED	60223	per Annum	672	21519	0	0	1940
18	8	4257	52	555	2020	464	2/22/2002	ACTIVE	80516	per Annum	260	79132	0	0	0
19	8	4257	52	555	2020	464	2/22/2002	SEASONAL	37.92	per Day	0	0	0	0	2222
20	2	4525	22	278	2020	902	5/20/2019	ACTIVE	44083	per Annum	1820	43313	7	169	262
21	5	4824	56	513	2020	467	7/2/2018	ACTIVE	78144	per Annum	1820	78227	0	0	0
22	5	4953	56	513	2020	467	1/1/2019	ACTIVE	94542	per Annum	1820	92514	0	0	0
23	2	5040	22	278	2020	902	3/2/2020	ACTIVE	44083	per Annum	560	13490	0	0	0
24	2	5144	22	278	2020	902	12/2/2019	ACTIVE	44083	per Annum	1015	23728	0	0	58
25	3	5171	68	516	2020	841	4/3/2006	ACTIVE	380.64	per Day	1976.15	93858	166	12085	4256
26	2	5296	22	278	2020	902	9/10/2012	ACTIVE	44083	per Annum	1820	44116	19	576	352
27	5	5437	62	2	2020	740	4/1/2012	ACTIVE	57921	per Annum	1820	57004	0	0	5766

Agencies Table, Personnel Table, Titles Table and Boroughs Table

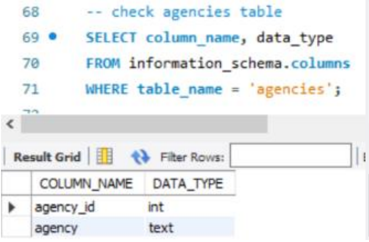
All the numeric values are in ID columns, and all other columns have correct text values. The middle initial column in Personnel Table has some missing values but no corrections are needed at this point.

1. Data type testing:



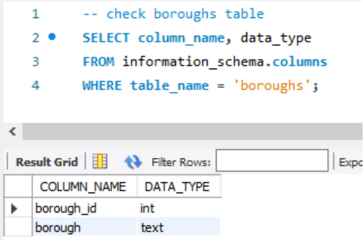
```sql
-- check titles table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'titles';
```

COLUMN_NAME	DATA_TYPE
title_id	int
title	longtext



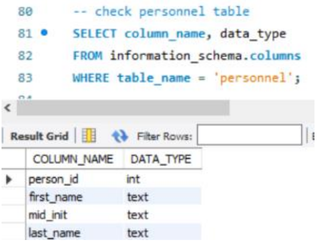
```sql
-- check agencies table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'agencies';
```

COLUMN_NAME	DATA_TYPE
agency_id	int
agency	text



```sql
-- check boroughs table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'boroughs';
```

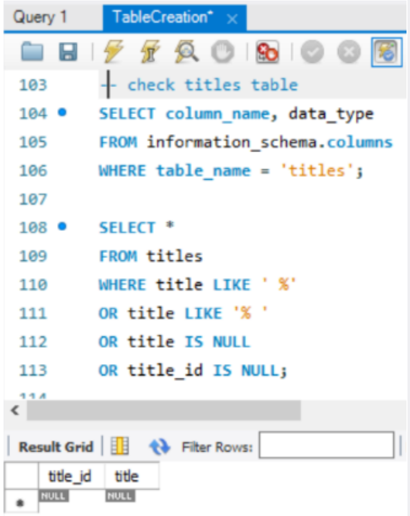
COLUMN_NAME	DATA_TYPE
borough_id	int
borough	text



```sql
-- check personnel table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'personnel';
```

COLUMN_NAME	DATA_TYPE
person_id	int
first_name	text
mid_init	text
last_name	text

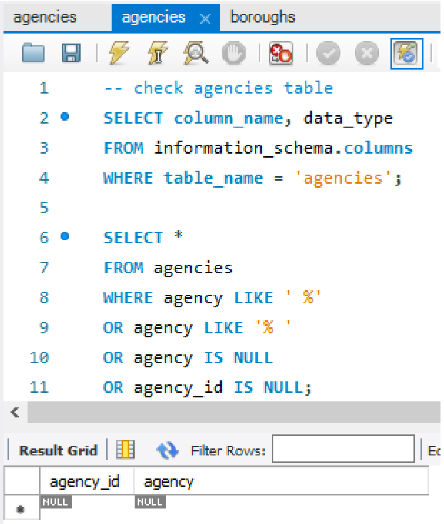
2. Table value testing:



```
-- check titles table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'titles';

SELECT *
FROM titles
WHERE title LIKE ' %'
OR title LIKE '% '
OR title IS NULL
OR title_id IS NULL;
```

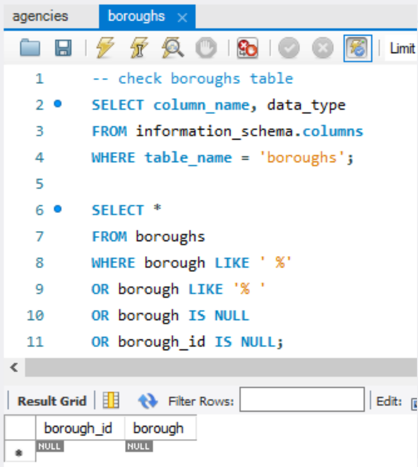
```
  title_id  title
* NULL      NULL
```



```
-- check agencies table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'agencies';

SELECT *
FROM agencies
WHERE agency LIKE ' %'
OR agency LIKE '% '
OR agency IS NULL
OR agency_id IS NULL;
```

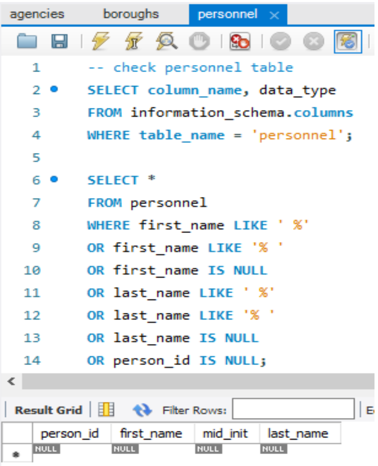
```
  agency_id  agency
* NULL       NULL
```



```
-- check boroughs table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'boroughs';

SELECT *
FROM boroughs
WHERE borough LIKE ' %'
OR borough LIKE '% '
OR borough IS NULL
OR borough_id IS NULL;
```

```
  borough_id  borough
* NULL        NULL
```



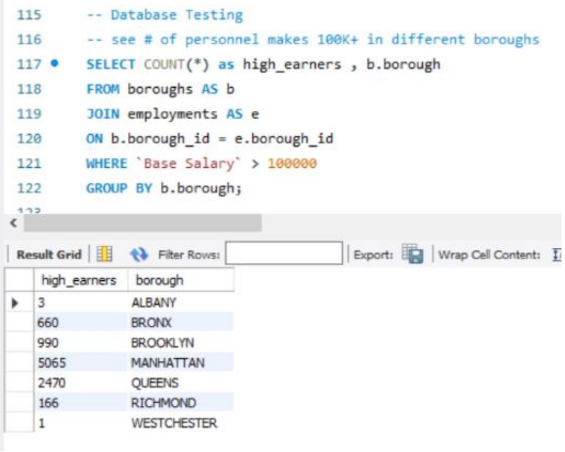
```
-- check personnel table
SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'personnel';

SELECT *
FROM personnel
WHERE first_name LIKE ' %'
OR first_name LIKE '% '
OR first_name IS NULL
OR last_name LIKE ' %'
OR last_name LIKE '% '
OR last_name IS NULL
OR person_id IS NULL;
```

```
  person_id  first_name  mid_init  last_name
* NULL       NULL        NULL      NULL
```

DATABASE TESTING

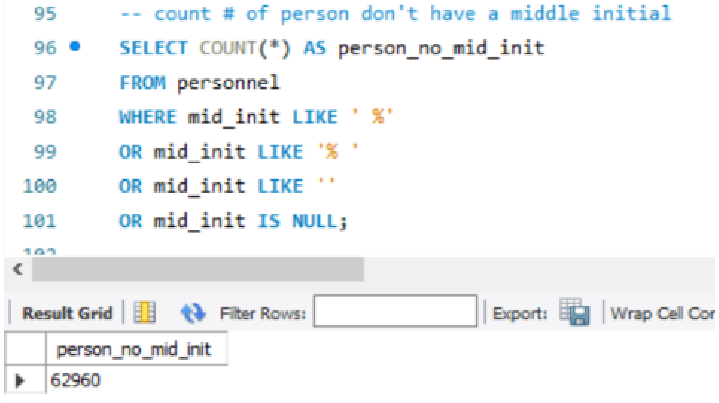
1. Check the number of employees who make more than $100,000 annualy in different boroughs



```
-- Database Testing
-- see # of personnel makes 100K+ in different boroughs
SELECT COUNT(*) as high_earners , b.borough
FROM boroughs AS b
JOIN employments AS e
ON b.borough_id = e.borough_id
WHERE `Base Salary` > 100000
GROUP BY b.borough;
```

```
high_earners  borough
3             ALBANY
660           BRONX
990           BROOKLYN
5065          MANHATTAN
2470          QUEENS
166           RICHMOND
1             WESTCHESTER
```

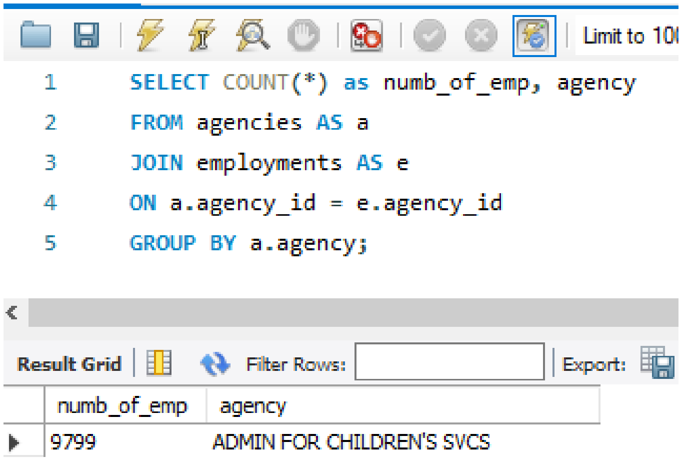
2. Check the average salary of people who don't have the middle name initials



```sql
-- count # of person don't have a middle initial
SELECT COUNT(*) AS person_no_mid_init
FROM personnel
WHERE mid_init LIKE ' %'
OR mid_init LIKE '% '
OR mid_init LIKE ''
OR mid_init IS NULL;
```

person_no_mid_init
62960

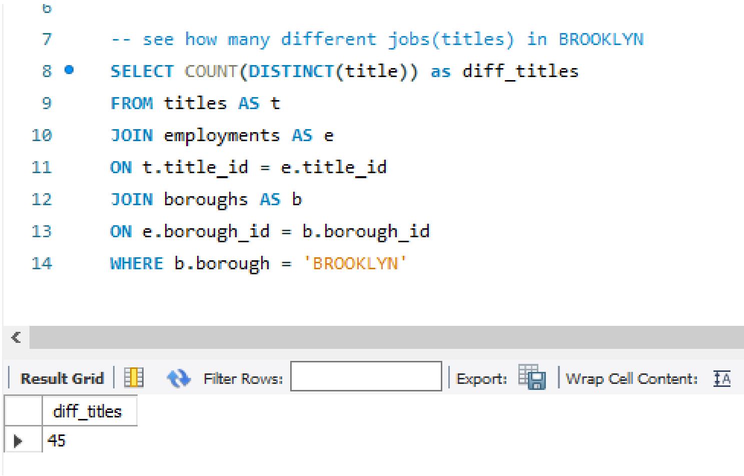
3. Check the distribution of employees through different agencies



```sql
SELECT COUNT(*) as numb_of_emp, agency
FROM agencies AS a
JOIN employments AS e
ON a.agency_id = e.agency_id
GROUP BY a.agency;
```

numb_of_emp	agency
9799	ADMIN FOR CHILDREN'S SVCS

4. Check how many different jobs(titles) in BROOKLYN



```sql
-- see how many different jobs(titles) in BROOKLYN
SELECT COUNT(DISTINCT(title)) as diff_titles
FROM titles AS t
JOIN employments AS e
ON t.title_id = e.title_id
JOIN boroughs AS b
ON e.borough_id = b.borough_id
WHERE b.borough = 'BROOKLYN'
```

diff_titles
45