CSC343 Project - Data Cleaning

1. Cleaning Process

- Dataset for Movie

  Since we're studying for movie ratings, we want our movie to be representative, so we only keep movies that have more than 600000 ratings (the top 2% of our original dataset) on IMDb.
  We are forcing the constraint "Myear <= current year" by removing the rows that don't satisfy this constraint since we don't want to include movies that do not release yet, the rating of those movies would not be what we want to study.
  We are forcing the "not null" constraint by removing the rows with null values since the null values would affect our queries for the question later, and we want to keep this dataset simple.

- Dataset for Finance

  In our original dataset, there are different currencies presented in it, so we convert all the values to US dollars so that we can compare the value correctly later.
  Since datasets for Movie and Finance are the same dataset, we don't need to do anything more for checking the foreign key constraint.
  We are forcing the "not null" constraint by removing the rows with null values since we do want to compare values between budget, USA revenue and global revenue later; still, we want to keep our dataset simple.

- Dataset for Rating

  We remove the rows whose Mid are not in the movie table since we are only considering those movies.
  We validate that the rating should be in the range 0 – 10 and remove the rows with invalid ratings.

- Dataset for Person

  We convert the birth dates to the same format (date type in SQL).
  We are forcing the "not null" constraint by removing the rows with a null value.
  We validate the birth date by comparing it to the current date.

- Dataset for Job

  We force the foreign key constraints by removing the rows that contain invalid data since we're only interested in movies from the movie table and stuff from the personal table.

- Dataset for Award

  We force the foreign key constraint by removing invalid rows since we only care about award records for the person in the person table.

2. Sample Data after Cleaning

- Sample Data for Movie

mid	title	myear	genre	duration	country	language
tt0012349	The Kid	1921	Comedy, Drama, Family	68	USA	English, None
tt0013442	Nosferatu, eine Symphonie des Grauens	1922	Fantasy, Horror	94	Germany	German
tt0015864	The Gold Rush	1925	Adventure, Comedy, Drama	95	USA	English, None
tt0017925	The General	1926	Action, Adventure, Comedy	67	USA	English
tt0017136	Metropolis	1927	Drama, Sci-Fi	153	Germany	German
tt0021749	City Lights	1931	Comedy, Drama, Romance	87	USA	English
tt0021884	Frankenstein	1931	Drama, Horror, Romance	70	USA	English, Latin
tt0022100	M - Eine Stadt sucht einen Mvərder	1931	Crime, Mystery, Thriller	117	Germany	German
tt0024216	King Kong	1933	Adventure, Horror, Sci-Fi	100	USA	English
tt0025316	It Happened One Night	1934	Comedy, Romance	105	USA	English

- Sample Data for Finance

mid	budget	usarevenue	revenue
tt0154506	6000	48482	48482
tt0390384	7000	424760	545436
tt0104815	7000	2040920	2040920
tt1179904	15000	107918810	193355800
tt0109445	27000	3151130	3151130
tt2866360	50000	102617	139745
tt0185937	60000	140539099	248639099
tt0138704	60000	3221152	3221152
tt1549572	100000	1321194	1938783
tt0063350	114000	236452	236452

- Sample Data for Rating

mid	allage	under18	from18to30	from30to45	above45
tt0012349	8.2	8.5	8.4	8.3	8.3
tt0013442	7.9	8.2	7.8	7.9	8.1
tt0015864	8.1	8	8.2	8.1	8.2
tt0017136	8.1	8.3	8.3	8.2	8.3
tt0017925	8	8	8.1	8.1	8.2
tt0021749	8.3	8.5	8.6	8.5	8.5
tt0021884	7.8	7.7	7.7	7.7	8.1
tt0022100	8.1	8	8.3	8.3	8.4
tt0024216	7.8	7.9	7.8	7.8	8.1
tt0025316	8	7.9	8.1	8	8.1

- Sample Data for Person

pid	name	birthdate	birthplace
nm0000001	Fred Astaire	1899-05-10	Omaha, Nebraska, USA
nm0000002	Lauren Bacall	1924-09-16	The Bronx, New York City, New York, USA
nm0000003	Brigitte Bardot	1934-09-28	Paris, France
nm0000004	John Belushi	1949-01-24	Chicago, Illinois, USA
nm0000005	Ingmar Bergman	1918-07-14	Uppsala, Uppsala l√§n, Sweden
nm0000006	Ingrid Bergman	1915-08-29	Stockholm, Sweden
nm0000007	Humphrey Bogart	1899-12-25	New York City, New York, USA
nm0000008	Marlon Brando	1924-04-03	Omaha, Nebraska, USA
nm0000009	Richard Burton	1925-11-10	Pontrhydyfen, Wales, UK
nm0000010	James Cagney	1899-07-17	New York City, New York, USA

- Sample Data for Job

mid	pid	jobcategory
tt0012349	nm0000122	actor
tt0012349	nm0701012	actress
tt0012349	nm0001067	actor
tt0012349	nm0588033	actor
tt0012349	nm0042317	actor
tt0012349	nm0074788	actor
tt0012349	nm0080930	actor
tt0012349	nm0088471	actor
tt0013442	nm0003638	director
tt0013442	nm0002302	composer

- Sample Data for Award

pid	awardname	ayear	outcome
nm1265067	Grammy	2008	Nominated
nm1265067	Grammy	2010	Won
nm1265067	Satellite Award	2005	Nominated
nm1265067	Grammy	2006	Nominated
nm1265067	Grammy	2004	Nominated
nm1265067	Black Reel	2004	Nominated
nm1265067	VMA	2005	Nominated
nm1265067	Teen Choice Award	2006	Nominated
nm0958989	WGA Award (Screen)	1970	Nominated
nm0958989	Screencraft Award	2006	Won