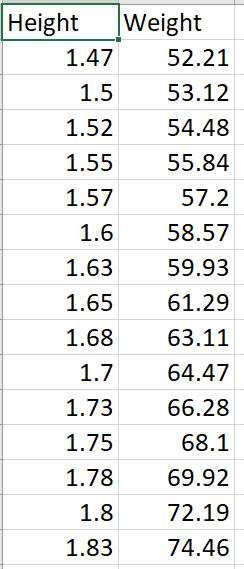
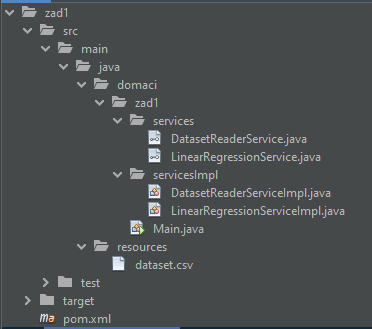
1. Domaći zadatak – Linearna regresija

**Uvod**: Cilj same regresione analize je predviđanje vrednosti zavisne promenljive na osnovu vrednosti nezavisne promenljive (promenljivih). Koraci u regresionoj analizi podrazumevaju: 1) Analizu zavisnosti koja važi između zavisne promenljive i nezavisne promenljive (ili nezavisnih promenljivih); 2) Kreiranje funkcije koja tu zavisnost opisuje – regresiona funkcija ili regresija.

Za potrebe ovog domaćeg zadatka korišćen je sledeći skup podataka (u .csv formatu), preuzet sa sajta <https://www.kaggle.com/> koji predstavlja zavisnost težine od visine:



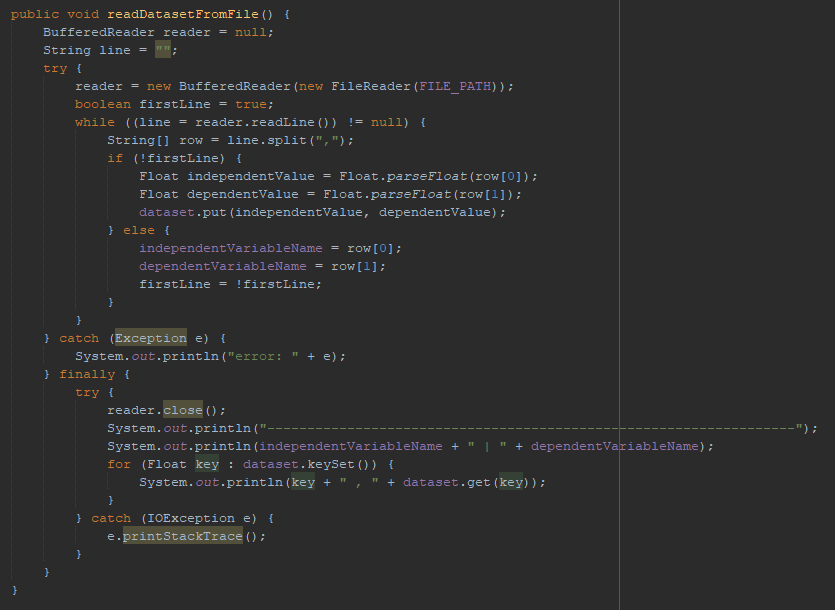
Multiplikativnost zavisnosti u prethodnom skupu podataka je jedan. Struktura samog projekta je prikazana na sledećoj slici:



Opis projekta:

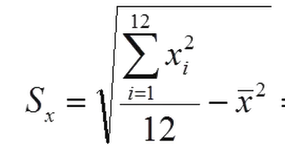
- **DatasetReaderService** je servis korišćen za čitanje skupa podataka iz „dataset.csv“ fajla koji se nalazi u resursima projekta. Njegov interface sadrži sledeće metode:

* **readDatasetFromFile**() – za učitavanje dataset.csv fajla, učitavanja imena nezavisne i zavisne promenljive, kao i popune mape(key,value) za pročitanje podatke (comma separated columns) iz fajla; vrši i logovanje učitanog dataset-a
* **getIndependentVariableName**() – za dovlačenje naziva nezavisne promenljive
* **getDependentVariableName**() – za dovlačenje naziva zavisne promenljive
* **getDataset**() – vraća skup podataka učitan iz fajla preslikan u mapu

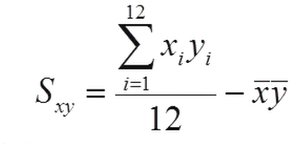


- **LinearRegressionService** je servis korišćen prevashodno za izračunavanje koeficijenata linearne regresije, ali i za računanje srednjih vrednosti, standardnih devijacija, kovarijanse promenljivih, korelacionog koeficijenta... Njegov interfejs sadrži sledeće metode:

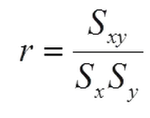
* calculateMean(List<Float> list) – za izračunavanje srednje vrednosti prosleđene liste vrednosti, koristi calculateMean(List<Float> list, boolean isSquareMean) da specificira da li se zahteva srednja vrednost ili kvadrat srednje vrednosti
* calculateStandardDeviation(List<Float> list) – koristi prethodnu metodu za računanje standarnih devijacija nezavisne i zavisne promenljive na osnovu prosledjene liste, po formuli



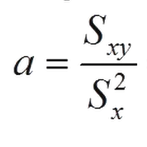
* calculateCovariance(List<Float> independentValues, List<Float> dependentValues) – za računanje kovarijanse promenljivih po sledećoj formuli



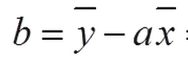
* calculateCorellationCoefficient(List<Float> independentValues, List<Float> dependentValues) – za računanje koeficijenta korelacije koji predstavlja reprezent jačine linearne zavisnosti, po formuli



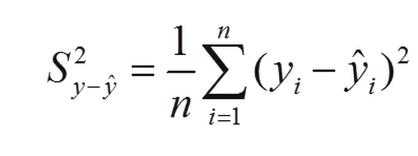
* calcualteIndependentCoefficientOfLREquotion(List<Float> independentValues, List<Float> dependentValues) – za računanje koeficijenta koji ide uz nezavisnu promenljivu u jednačini linearne regresije, po formuli



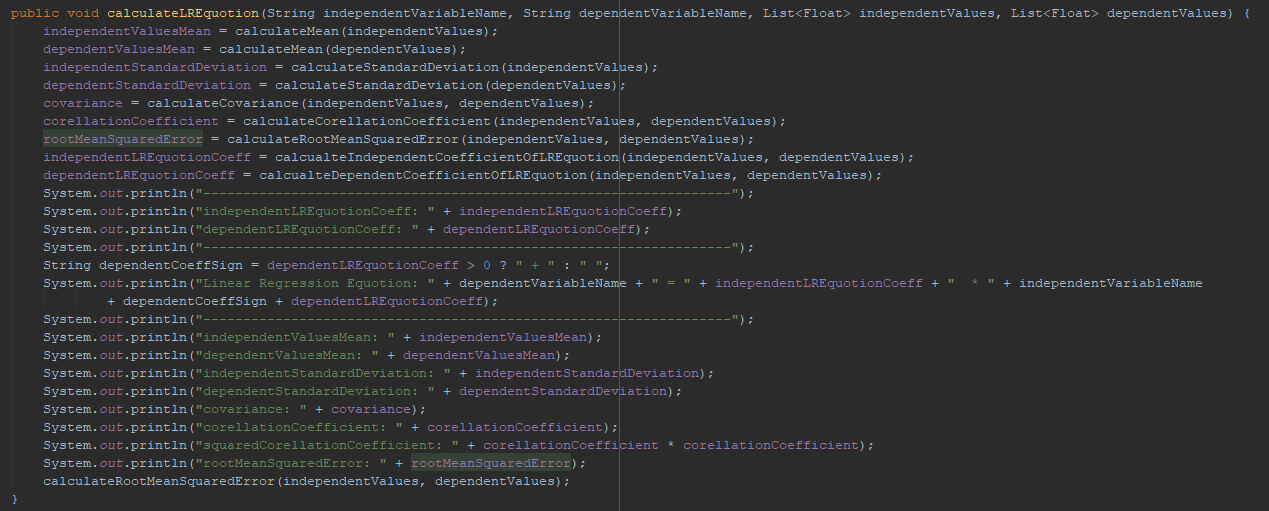
* calcualteDependentCoefficientOfLREquotion(List<Float> independentValues, List<Float> dependentValues) - za računanje koeficijenta koji ide uz zavisnu promenljivu u jednačini linearne regresije, po formuli



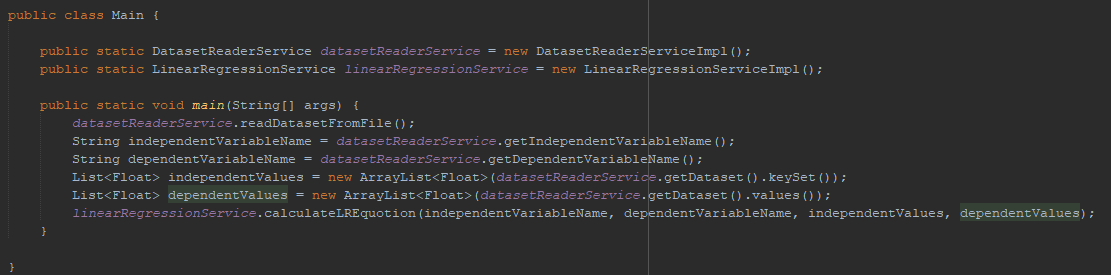
* calculateRootMeanSquaredError(List<Float> independentValues, List<Float> dependentValues) – za računanje standardne devijacije u odnosu na procenjenu vrednost (za razliku od klasične standardne devijacije koja se računala u odnosu na srednju vrednost, ovde računamo u odnosu na procenjenu vrednost), po formuli



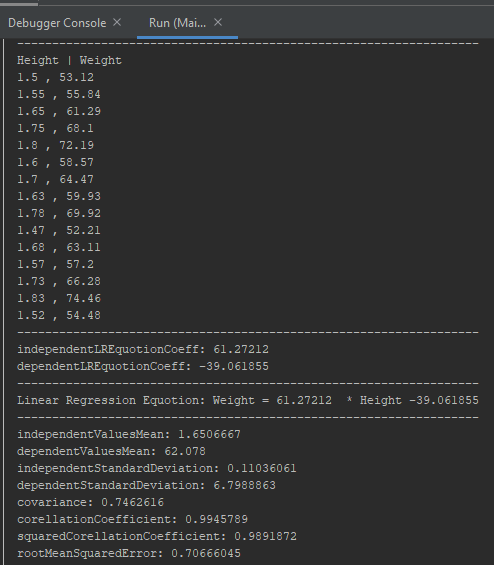
* calculateLREquotion(String independentVariableName, String dependentVariableName, List<Float> independentValues, List<Float> dependentValues) – za izračunavanje i logovanje jednačine linearne regresije i prethodno pomenutih parametara



**Glavni program** – inject-uje prethodno pomenute servise (datasetReaderService koristi za čitanje skupa podataka iz „dataset.csv“ fajla koji se nalazi u resursima projekta, a linearRegressionService za izračunavanje jednačine linearne regresije, pratećih parametara i međurezultata, kao i mera kvaliteta predviđanja)



**Rezultat** **koji program loguje**:



Dobijeni koeficijenti linearne regresije su: 61.27212 i -39.061855

Na osnovu kvadrirane vrednosti korelacionog koeficijenta (0.9891872) sa slike se zaključuje da imamo jaku linearnu zavisnost za izabrani skup podataka (odgovara opsegu jake linearne zavisnosti), pa se u ovom slučaju može vršiti sama regresija, tj. predviđanje.

Na slici je takođe prikazana i standardne devijacija u odnosu na procenjenu vrednost (0.70666045).

**Vizuelni prikaz:**

