Biomeetria praks 6

Illustreeritud (mittetäielik) tööjuhend

Eeltöö

1. Avage *MS Excel*'is oma kursuse ankeedivastuseid sisaldav andmestik,
2. lisage uus tööleht, nimetage see ümber leheküljeks 'Praks6' ja
3. kopeerige kogu 'Andmed'-lehel paiknev andmetabel lehekülje 'Praks6' ülemisse vasakusse nurka.

Ülesanne 1.

- Illustreerige tunnuste 'PIKKUS' ja 'JALANR' vahelist seost hajuvus- ehk punktdiagrammiga.
- Jälgige, et x-telg (horisontaalne telg) vastaks jalanumbritele ja y-telg (vertikaalne telg) pikkustele. Vajadusel kujundage joonis ümber.
- Prognoosimaks pikkust jalanumbri alusel, lisage joonisele lineaarne regressioonisirge, samuti regressioonivõrrand ja viimase alusel leitavate prognooside täpsust kirjeldav determinatsioonikordaja R^2.
- Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng.

Ülesanne 2.

- Teostage statistikaprotseduuri `Regression` (`Data`-sakk -> `Data analysis`…) abil lineaarne regressioonanalüüs prognoosimaks tudengite pikkust jalanumbri alusel.
- Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul

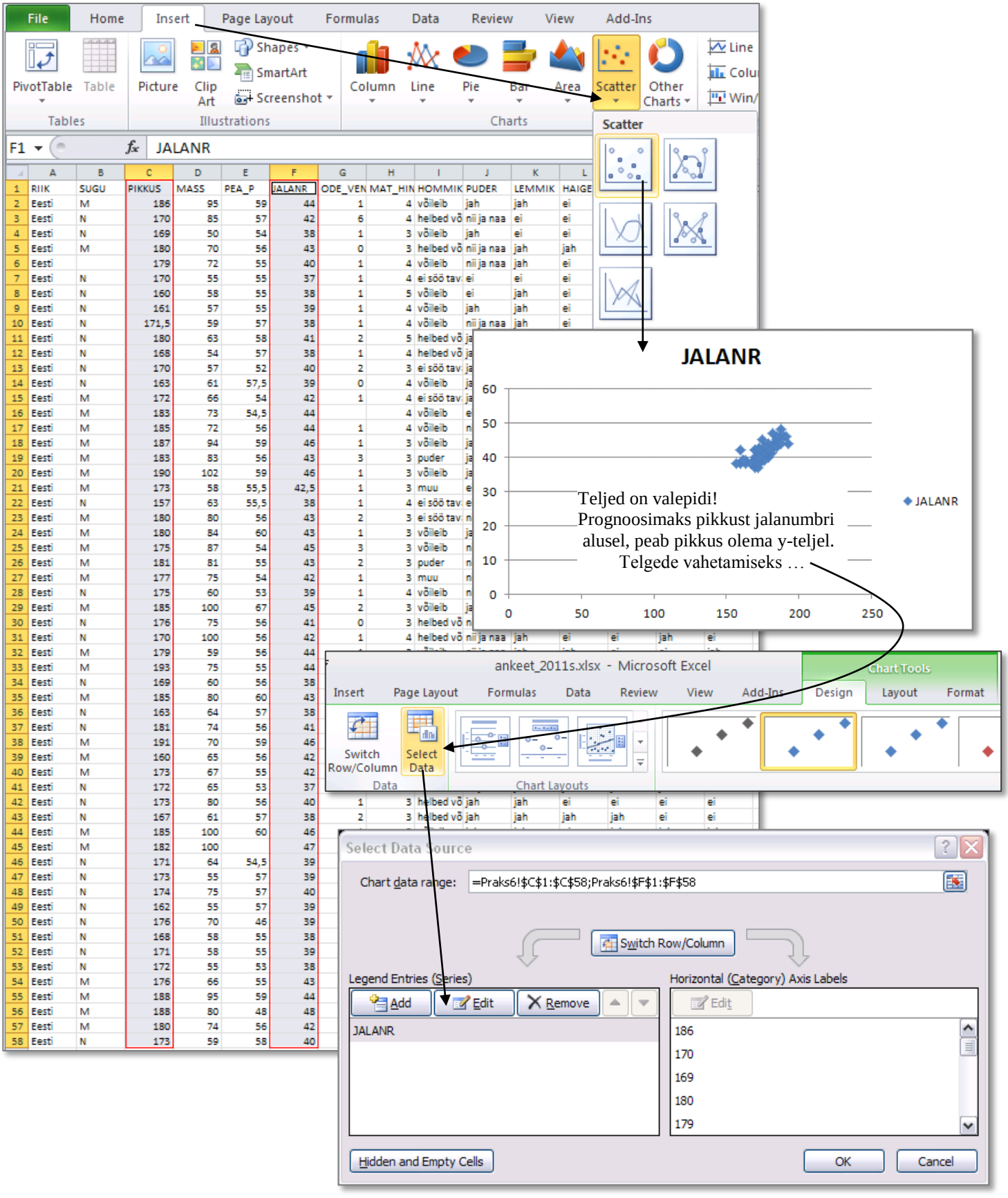
  $$Pikkus = a + b \times Jalanumber,$$

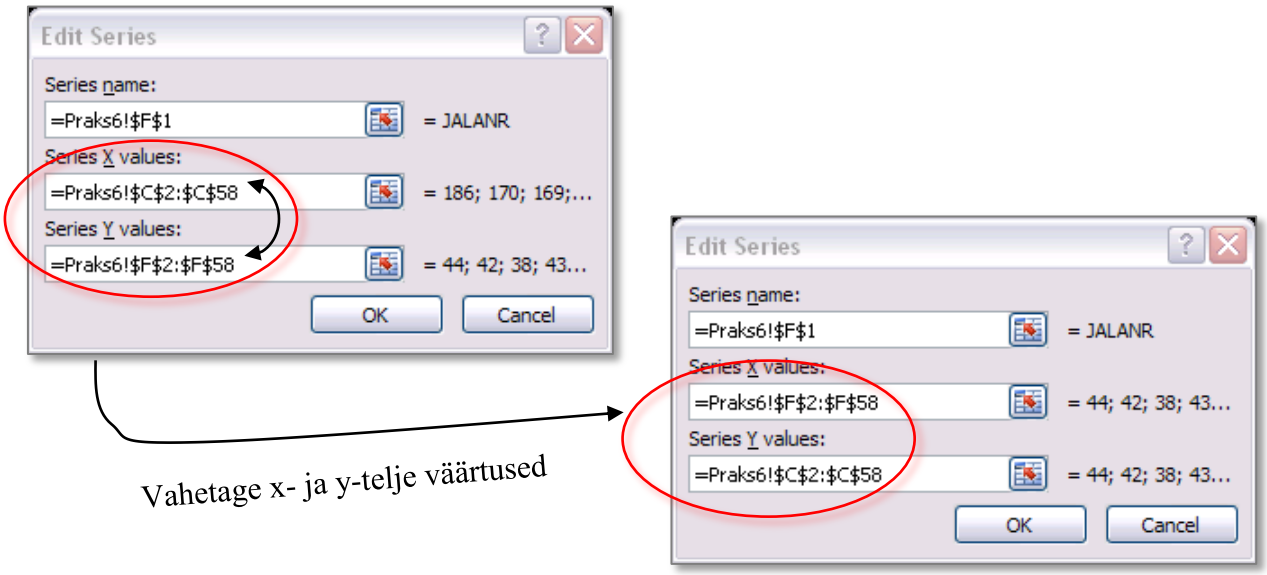
  kus *a* ja *b* asemel on *Excel*'i poolt välja arvutatud kordajate väärtused.
- Kui suur on keskmiselt pikkuste vaheline erinevus tudengitel, kelle jalanumbrid erinevad 2 võrra?
- Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!
- Kirjeldamaks prognooside täpsust, sõnastage üks lause kas mitmese korrelatsioonikordaja (*R*), mudeli standardvea (*Standard Error*) või determinatsioonikordaja (R^2) kohta.

Ülesande 1 tööjuhend

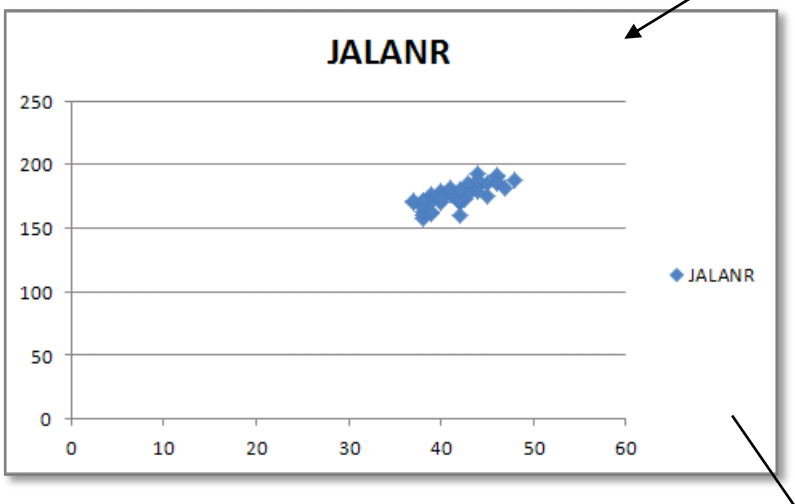
1. Illustreerige tunnuste ’PIKKUS’ ja ’JALANR’ vahelist seost hajuvus- ehk punktdiagrammiga.

   Joonisel peab x-telg vastama jalanumbritele ja y-telg pikkustele. Vajadusel kujundage joonis ümber (vt allpool toodud juhendit)**.**

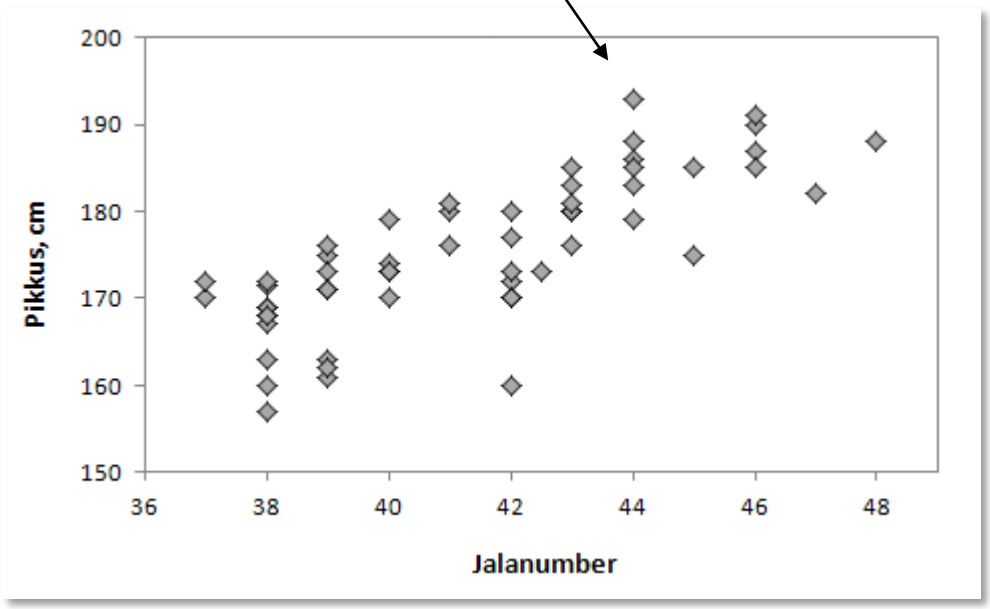




Tulemus:
pikkuse väärtused on y- ja
jalanumbri väärtused x-teljel.

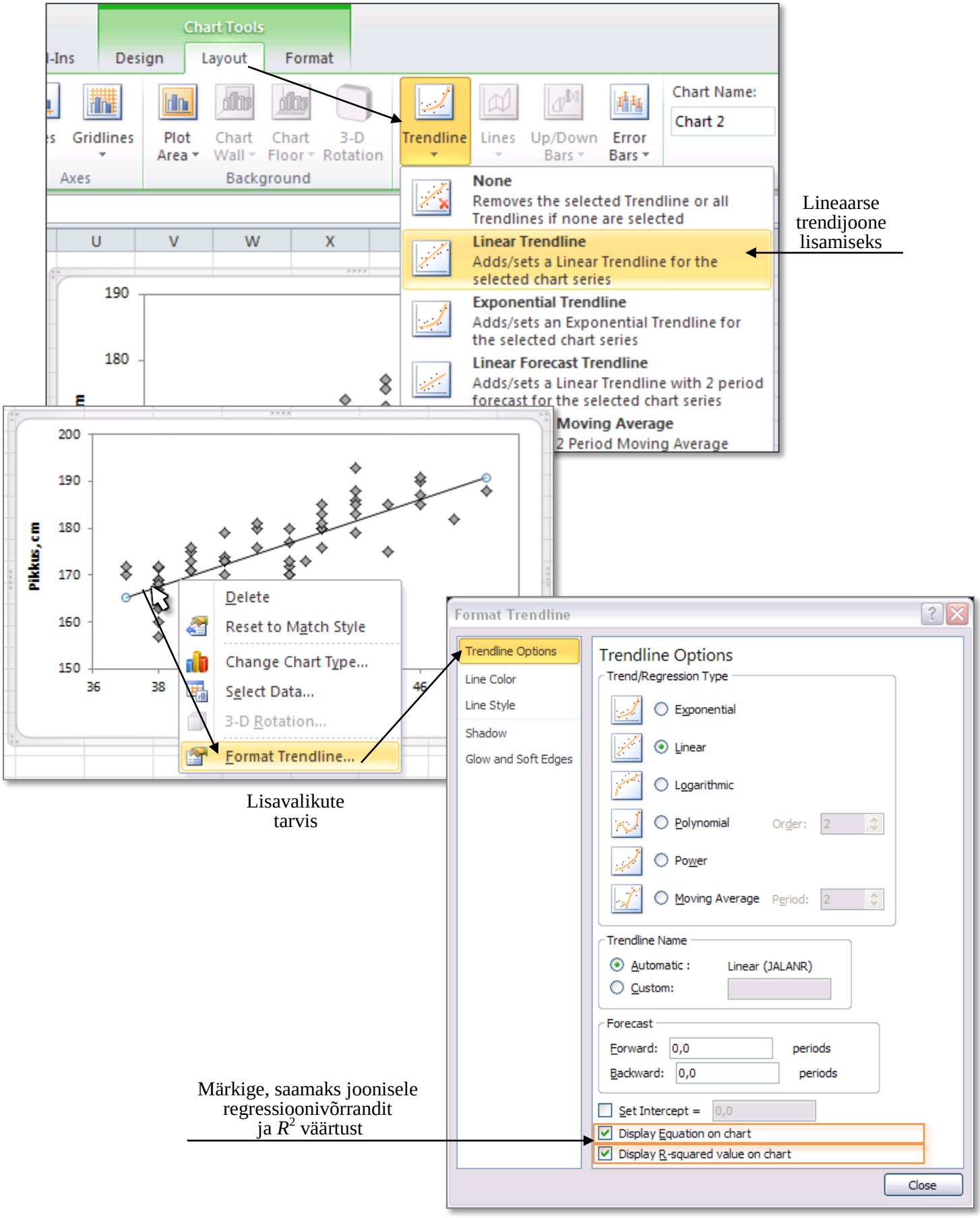


Kujundage joonis!



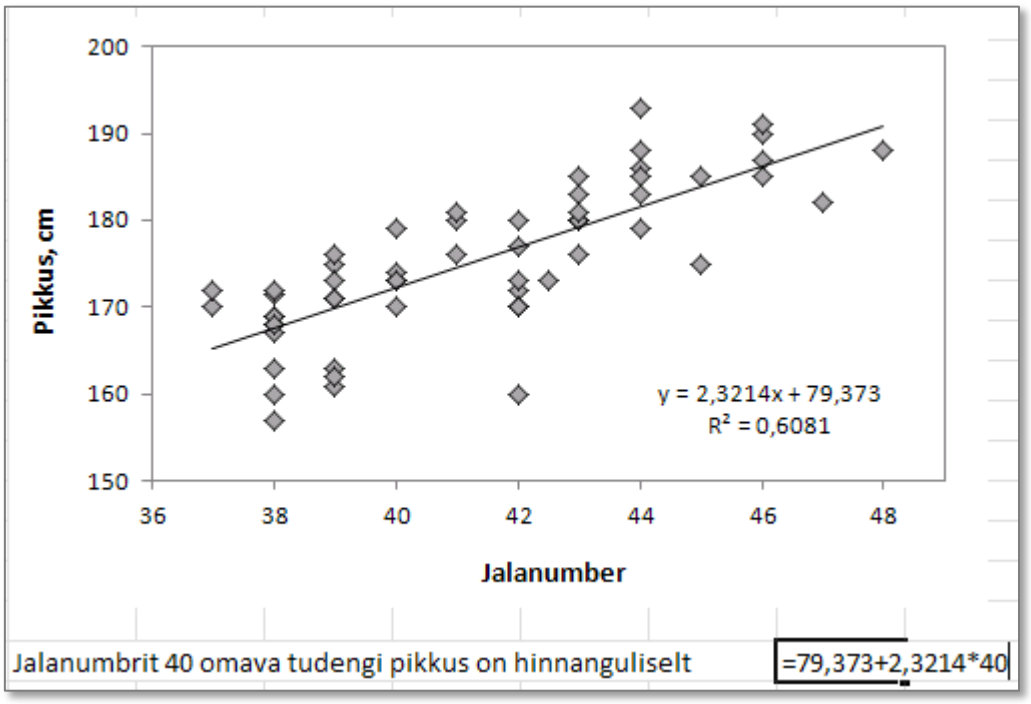
2. Prognoosimaks pikkust jalanumbri alusel, lisage tunnuste ’PIKKUS’ ja ’JALANR’ hajuvusdiagrammile **regressioonisirge**.

Lisage joonisele ka **regressioonivõrrand** ja viimase alusel leitavate prognooside täpsust kirjeldav **determinatsioonikordaja** R^2.



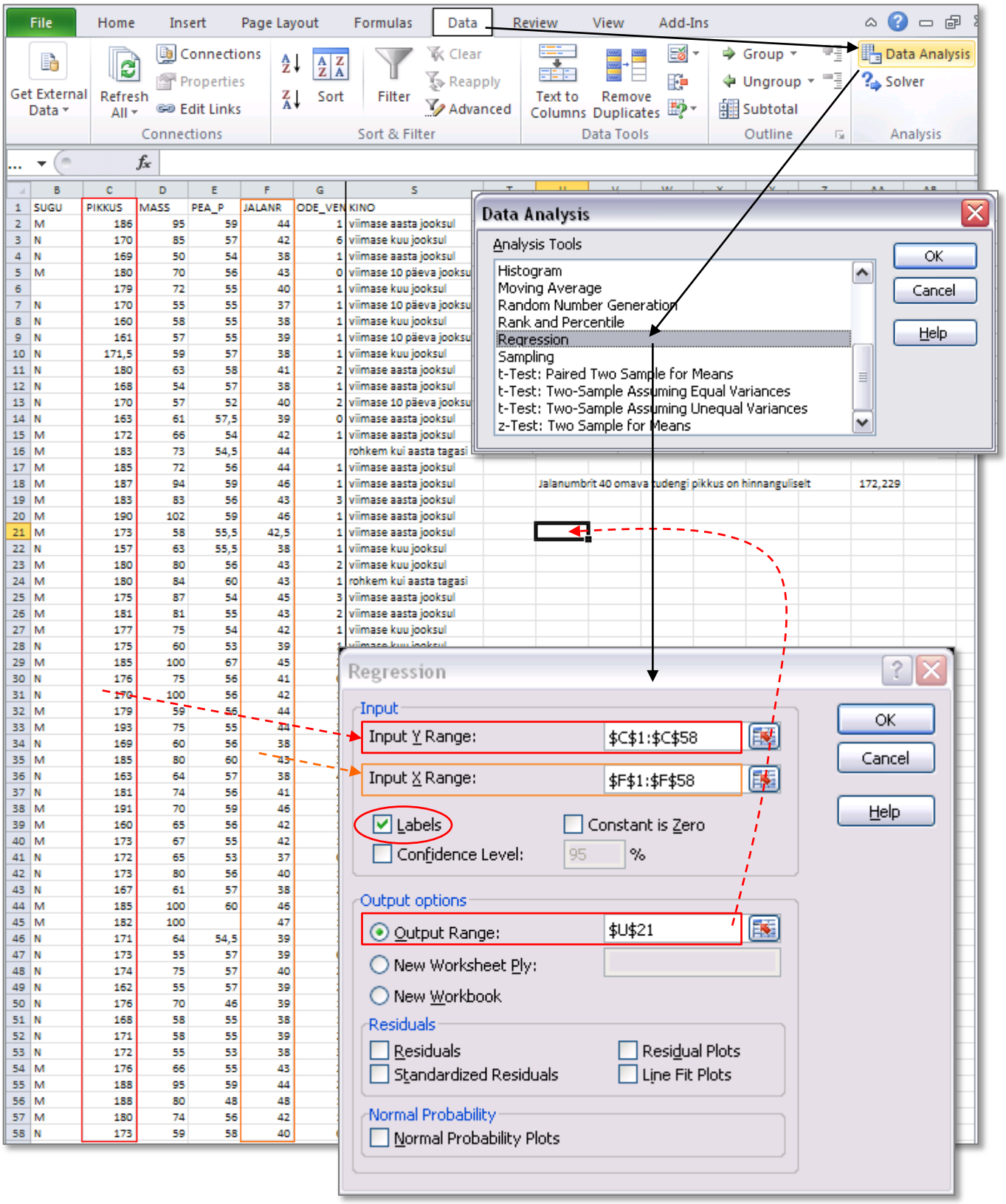
3. Prognoosige leitud võrrandi alusel, keskmiselt kui pikk on jalanumbrit 40 omav tudeng.

Selleks pange joonise alla kirja *Excel*'i poolt välja arvutatud regressioonivõrrand, asendades lihtsalt suuruse *x* arvuga 40. ☺



Ülesande 2 tööjuhend.

1. Teostage statistikaprotseduuri `Regression` (`Data`-sakk -> `Data analysis`…) abil lineaarne regressioonanalüüs prognoosimaks tudengite pikkust jalanumbri alusel.



Regressioonanalüüsi tulemus:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,77980172
R Square	0,60809072
Adjusted R Square	0,60096509
Standard Error	5,37994256
Observations	57

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2470,02182	2470,022	85,3386	8,82306E-13
Residual	55	1591,908005	28,94378		
Total	56	4061,929825			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	79,3727035	10,42634534	7,612706	3,74E-10	58,47784053	100,26757
JALANR	2,32136296	0,251287157	9,237889	8,82E-13	1,817772241	2,8249537

2. Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul

$$Pikkus = a + b \times Jalanumber,$$

kus *a* ja *b* asemel on *Excel*'i poolt välja arvutatud kordajate väärtused.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	79,3727035	10,42634534	7,612706	3,74E-10	58,47784053	100,26757
JALANR	2,32136296	0,251287157	9,237889	8,82E-13	1,817772241	2,8249537

3. Kui suur on keskmiselt pikkuste vaheline erinevus tudengitel, kelle jalanumbrid erinevad 2 võrra?

Vastus: 2×*b* (aga arvuliselt?). **Pange arvuline vastus kirja täislausega.**

4. Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2470,02182	2470,022	85,3386	8,82306E-13
Residual	55	1591,908005	28,94378		
Total	56	4061,929825			

$= p$

Märkus. Regressioonivõrrandi statistiline olulisus tähendab seda, et leitud regressioonivõrrand kujul

$$Pikkus = a + b \times Jalanumber$$

võimaldab pikkust täpsemalt prognoosida võrreldes konstantse võrrandiga

$$Pikkus = a.$$

Ehk siis, statistiliselt olulise regressioonivõrrandi korral võimaldab jalanumbri arvestamine pikkust täpsemalt prognoosida võrreldes konstateeringuga, et kõigi tudengite pikkused on ühesugused (ja võrdsed suurusega a).

Hüpoteeside paar, mille testimiseks vajaliku p-väärtuse väljastab *Excel* tabelisse ANOVA, on kujul:

H_0: regressioonivõrrand ei ole statistiliselt oluline
H_1: regressioonivõrrand on statistiliselt oluline

ehk

H_0: leitud võrrand ei ole parem võrreldes konstantse võrrandiga
H_1: leitud võrrand on parem võrreldes konstantse võrrandiga

ehk matemaatiliselt

H_0: $Pikkus = a$
H_1: $Pikkus = a + b \times Jalanumber$

Reaalselt rakendada on põhjust vaid statistiliselt olulist regressioonivõrrandit.

5. Sõnastage üks lause regressioonivõrrandist saadavate prognooside täpsuse kohta kas mitmese korrelatsioonikordaja (R), determinatsioonikordaja (R^2) või mudeli standardvea baasil.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,77980172
R Square	0,60809072
Adjusted R Square	0,60096509
Standard Error	5,37994256
Observations	57

Mitmene korrelatsioonikordaja R mõõdab uuritava tunnuse ja tema prognoositud väärtuste vahelist korrelatsiooni. Mida suurem, seda parem!

Determinatsioonikordaja R^2 näitab, kui suure osa uuritava tunnuse varieeruvusest võrrandist saadud prognoosid ära kirjeldavad, $0 \le R^2 \le 1$. Esitatakse enamasti protsentides. Mida suurem, seda parem!

Mudeli standardviga SE on prognoosijääkide standardhälve. Näitab tegelike ja prognoositud väärtuste vahelist keskmist erinevust (mudeli keskmist viga). Mida väiksem, seda parem!