Topics in Data Science = Өгөгдлийн шинжлэх ухаан

Data Visualization – Өгөгдлийн визуальчлал

Монгол Улсын Үндэсний статистикийн хороо, Улаанбаатар хот, Монгол Улс, 2017 оны 5 сарын 8

Ryan Womack

Data Librarian, Rutgers University, https://ryanwomack.com

Хэсэг 1: Танилцуулга

Илтгэгчийн танилцуулга

• Data Librarian

• Эдийн засаг, номын сангийн шинжлэх ухаан, статистикийн магистрийн зэрэгтэй

- Эдгээр салбарын огтлолцол дээр ажилладаг
- Өгөгдөл хайх, ашиглахад нь хүмүүст тусалдаг
- Номын санд зориулан мэдээллийг удирддаг

• Ахисан түвшний эрдэм шинжилгээ, судалгаан дээр ажиллах

Rutgers-ийн талаарх танилцуулга

• Rutgers, The State University of New Jersey

• New Jersey, 8.8 сая хүн амтай, New York болон Philadelphia-ийн хооронд оршдог

• Rutgers 1766 онд байгуулагдсан, 250 гаруй жилийн түүхтэй

• Carnegie ангилал: Судалгаа – Судалгааны ажлыг маш өндөр түвшинд хийдэг, дээд түвшний ангилал (R1)

• 100 гаран судалгааны хөтөлбөртэй

• 68,000 гаруй оюутантай (120 гаруй улсын 7,500 олон улсын оюутантай)

• Дэлхийн топ 100 их сургуулийн нэг (Times Higher Education, Shanghai Ranking, CWUR болон бусад)

Rutgers-ийн талаарх танилцуулга

Зураг

IASSIST-ийн талаарх танилцуулга

• IASSIST – Нийгмийн шинжлэх ухааны мэдээллийн үйлчилгээ, технологийн олон улсын нийгэмлэг

• Мэдээллийн технологи, номын сан, өгөгдлийн үйлчилгээ, судалгаа болон дээд боловсрол, төрийн, ашгийн бус болон хувийн судалгааны салбарт ажиллаж буй 300 гаруй өгөгдлийн мэргэжилтнүүдийн нэгдэл

• IASSIST дараахь зорилгын хүрээнд ажилладаг

- Өгөгдлийн үйлчилгээг хүргэх төгс сүлжээг дэмжих сурталчлах
- Нийгмийн шинжлэх ухааны дэд бүтцийг сайжруулах

- Мэргэжлийн туршлага солилцох боломжийг бүрдүүлэх

• Жил бүрийн конференци нь харилцан туршлага судлах, харилцаа холбоо үүсгэх фоум болдог (АНУ, Канад, Европ, дараагийнх Азид?)

• IASSIST Ази дахь гишүүнчлэлээ нэмэгдүүлэхийг зорьж байгаа.

Яагаад өгөгдлийн визуальчлал гэж?

Өгөгдлийн визуалчлал нь:

• өгөгдлийн шинж чанарын талаар тодорхой ойлголт өгдөг

• өгөгдөл дэх нуугдмал бүтцийг илрүүлж өгдөг

• мэдээллийг хураангуйлж өгдөг

Anscombe's quartet буюу Анскомбегийн дөрвөл

Жишээ нь, Anscombe's quartet буюу Анскомбегийн дөрвөлийг харна уу

(зургийн эх сурвалж: http://commons.wikimedia.org/wiki/File:Anscombe%27s_quartet_3.svg):

Зураг:

Хэсэг 2: Түүх

Playfair

• Астраном (Одон орон судлал)-ын ажиглалт, график, газрын зураг 1800 оноос өмнө график дүрслэлд шинэчлэлийг авчирчээ. Сонгодог Өгөгдлийн Визуалчлал (Classic Data Visualizations)-ыг мөн үзнэ үү

• William Playfair бол шугаман график (line chart), дөрвөлжин дүрслэл бүхий гарфик (bar chart), цаг хугацааны график (time series plots), дугай график (pie chart)-ийг анх гаргасан хүн.

• Playfair, W. (1786). Худалдааны болон улс төрийн атлас: зэс тавагны дүрслэл (Copper-Plate Charts)-ээр 19 дүгээр зууныг бүхэлд нь Англи улсын худалдаан, орлого, зарлага, өрийн хэмжээний өсөлтийг харуулж байсан,

• Playfair, W. (1801). Statistical Breviary (Статистикийн товчоон).

• Худалдааны болон улс төрийн атлас болон Статистикийн товчооныг хоёуланг нь Кэмбриджийн их сургуулийн хэвлэх газраас 2005 онд дахин хэвлэн гаргасан.

Playfair-ийн жишээ

Зураг:

Playfair-ийн жишээ

Зураг:

Minard

Charles Joseph Minard бол Playfair-ийн дараахь өгөгдлийн графикийг бий болгосон гол төлөөлөгч байсан.

• Minard-ийн Наполеоны Орос руу хийсэн аян дайны зураглалыг Tufte болон бусад хүмүүс мэдээллийг график дүрслэлд оруулсан агуу бүтээлүүдийн нэг гэж үзсэн байдаг.

• Энд мэдээллийн элементүүдийг хамгийн өндөр түвшинд нэгтгэн харуулж чадсан байдаг

• Зургаан төрлийн хувьсагч: 2 хэмжээсээр хэмжээ болон байршил, армийн чиглэл, температур, огноо [болон бүлэг]

• Гэхдээ, энэ нь инфографикийн нэг хэлбэр боловч үүнийг R болон бусад програм хангамжаар хийж болно.

Minard-ийн жишээ

Зураг:

Minard-ийн жишээ

Зураг:

Fisher болон Tukey

• 20 дугаар зуунд Ronald Fisher болон John Tukey зэрэг статистикчид өгөгдлийн дүн шинжилгээнд графикийн аргыг ашиглах арга зүйг үргэжлүүлэн хөгжүүлсэн.

• Fisher хамаарлыг ойлгоход өгөгдлийг хэсэгчлэн дүрслэх (plotting the data) аргыг тодорхойлсон байдаг.

• Tukey-ийн Тайлбарлах өгөгдлийн шинжилгээ (Exploratory Data Analysis) гаднах уншигчдад эцийн үр дүнг танилцуулахаас илүүтэй өгөгдлийг ойлгоход график дүрслэлийг хэрэглэхийг онцолсон байдаг.

• Tukey box болон whiskers plot, stem болон leaf plot-ийг бий болгосон.

Хэсэг 3: Визуалчлалын хольц

Bar chart-аас Dot plot

• Cleveland dot plot

– эрэмбэлсэн жагсаалтаар нэрлэсэн тоо хэмжээг харьцуулахад ашигладаг

Зураг:

Barchart болон Dot Plot

Өгөгдлийн тархалтыг визуал байдлаар харуулах

• Box болон Whiskers Plot

– Квантил (quantiles) болон хэт өндөр болон бага утгууд (outliers)-ыг харуулдаг. Мөн Tufte-ийн хувилбар байдаг.

• Violin plot

– Нягтралын мэдээллийг box болон whiskers хэлбэрээр хослуулан харуулдаг (урлагийн хэв маягаар)

Зураг:

Зураг:

Box Plot болон Violin Plot

Категорийн өгөгдлийг харуулах

• Pie chart-ын дараагийн хэлбэр

• Mosaic plot нь нэг графикт олон категорийг харуулах боломжийг олгодог, гэхдээ тайлбарлахад хүндрэлтэй.

• *Spineplot* бол a variant of the mosaic plot-ийн нэг хувилбар бөгөөд 2 хэмжээст харьцааг харуулдаг.

Зураг:

Зураг:

Pie Chart болон Mosaic Plot

Газрын зураг болон товгор дүрслэл (Glyph)

Газрын зураг нь өгөгдлийг харулах чухал арга бөгөөд өргөн хэрэглэгддэг.

• Өгөгдлийн түвшинг сүүдэрлэн харуулах choropleth газрын зургийг цөөн тооны жишээг бид ашигладаг.

• Мөн цаашдын судалгаандаа R дээрх Interactive Maps болон түүний 5 төрлийг авч үзнэ үү

Товгор дүрслэл (Glyphs) өгөгдлийн элементийн дүрсэн төлөөллийг харуулдаг.

• Цаг уурын газрын зурагт ихэвчлэн glyphs-ийг ашигладаг.

• Илүү динамик жишээ энд байна.

• R-ийн жишээн дээр Chernoff faces болон aplpack package-ийг авч үзнэ үү. Мөн, Smiley faces [энэ бүлэг дэх олон график дүрслэлийг хувилбар байгаа].

Зураг:

Зураг:

Choropleth Map болон Chernoff Faces

Хэсэг 4: Өгөгдлийн интерактив визуалчлал (Interactive Data Visualization)

Interactive DataViz - зарчмууд

• Яагаад бидний бүх график интерактив биш байна?

• Багсаар будах аргыг өгөгдлийг цэгүүдийг сонгох, тэдгээрийг янз бүрийн дүн шинжилгээ хийхэд ашигладаг.

• Салгаж авах, томруулах, жижиг хэсэгт хуваах нь мөн интерактив техник юм.

• Өгөгдлийн харуулах байдал нэг панелд сонгон авсан хэсэг нь өөр нэг панелд үр дүнг харуулах холбоостой байдлаар хийгдэж болно.

• Интерактив байдал (Interactivity) өгөгдлийг олж илрүүлэх, олон хэмжээст хамаарлыг судлахад ялангуяа хэрэглэгддэг.

Практик дээрх интерактив визуалчлал

График хэрэглэгчдийн интерфэйст интерактив өгөгдлийг ажиллуулах боломжийг олгох олон пакет R програмд байдаг:

• playwith – аливаа график функцтэй ажиллах өөрчлөгддөг пакет. Графикийг засварлаж, экспортлож болно.

– GTK+ -ийг компьютер дээрээ тусад суулгах шаардлагатай [OS-оор ялгагддаг арга]

googleVis

Ихэнх тохиолдолд өгөгдлийн элемент хоорондын хамаарлыг харуулахыг холбогдох өгөгдлийг интерактив байдлаар харуулах замаар хялбар шийдэж болно.

• Үүнийг googleVis болон бусад “Vis” пакетууд, ж.нь. биологийн төрөл зүйлийн хувьд bdvis эсвэл rainfreq зэргээр хялбарчилж болно.

• Номын сангийн жишээ - төрийн CIC их сургуулиудын хувьд сонгон авсан ARL Statistics-ийг харьцуулахад

Вэбэд суурилсан Интерактив өгөгдөл - Rcharts

• Rcharts бол интерактив визуалчлалыг хийхэд javascript-ийг ашигладаг пакет юм.

• Орон зайн тор загварын (Lattice-style) командуудыг ашигладаг.

• Уг пакетаар HTML page-д ашиглах javascript-ийг бий болгож болно.

• Зарим команд нь суулгасан байх ёстой NVD3 гэх мэт нэмэлт javascript-ийн архиваас хамаардаг

• slidify-ийн хамтаар мөн баримтуулалтад ашиглаж болно

Вэбэд суурилсан Интерактив өгөгдөл - shiny

• shiny пакетыг Rstudio-ийн хүмүүс хөгжүүлдэг

• Та онлайн хичээлээр дамжуулан shiny-г хагас өдөр сурч чадна.

• Дизайны илүү хэрэглэгчийн тодорхойлсон хяналтыг бусад do-it-all пакетуудтай харьцуулсан байдлаар shiny дээр хийх боломжтой

• Графикууд нь вэбийн ажиллагааг хэрэгжүүлэх wrapper-тай Graphics use familiar R-д танигдах syntax (ggplot2 зэрэг)-ийг ашигладаг

• shiny-гийн аппликэшн бүр нь ижил бүтэцтэй: R-ийн хоёр script [ui болон server files] нэг хавтсанд хамтдаа хадгалагдсан байдаг

• Вэбээр хуудсыг оруулахын тулд shiny серверийг суулгах шаардлагатай

Вэбэд суурилсан Интерактив өгөгдөл – shiny, үргэлжлэл

• shiny пакетийн бий болгосон загварууд байдаг.

• shiny-гийн галлерейгаас илүү ихийг харж болно

• Rcharts нь бас shiny-тай хамт ажилладаг.

Вэбэд суурилсан Интерактив өгөгдөл – ggvis

• ggvis пакетыг Rstudio-ийн хүмүүс мөн хөгжүүлдэг

• ggplot нь shiny-тай тохирдог гэдгийг санах

• ggplot-той синтакс нь нь адилхан

• Интерактив хяналтыг нэмэх зарим боломжийг агуулсан

• Вэбэд shiny-г ашиглаж болно

Вэб визуалчлалын бусад (R-аас бусад) хувилбарууд

• D3.js, http://d3js.org/ сайтаас үнэгүй татаж авч болно

• Inkscape, https://inkscape.org/ сайтаас үнэгүй татаж авч болно

• Tableau, оюутнуудад зориулсан нэг жилийн үнэгүй лицензийг
http://www.tableau.com/academic/students

• Plot.ly environment at http://plot.ly

• Datavisualizationforall – аргачлалуудтай үнэгүй онлайн ном

Интерактив хүч

• Хүн ам зүйн суварга interactivity+animation=insight –ийн нэг жишээ.

• Populationpyramid.net – бүх улс орнуудын хувьд, үндсэн анимэйшнтэй

• Destatis-ийн Герман улсын хүн ам зүйн суварга нь илүү интерактив хэлбэртэй

• Эдгээр загвар (Хэсэг 1) болон (Хэсэг 1)-ын дагуу R програм дээр хийх боломжтой

• ggvis пакетыг Rstudio-ийн хүмүүс мөн хөгжүүлдэг

Сүүлийн үеийн хөгжүүлэлтүүд болон хэрэглүүрүүд

• Дайн болон Энх тайвны талаарх сэтгэл хөдлөлийн аянд ойлголтыг төрүүлэхэд R-ийг ашиглах

• Hourly Heatmap

• ropensci – нээлттэй шинжлэх ухаанд зориулсан

• Digital Panopticon – зарим хэрэглүүрүүд

Хэсэг 5: Их хэмжээний өгөгдөл

Их хэмжээний өгөгдөл

• Их хэмжээний өгөгдөл нь өгөгдлийн визуалчлалын онцгой асуудлын нэг юм

• Олонх техник, графикууд нь адилхан хэдий ч өгөгдлийн сангийн хэмжээний тохирсон аргыг хэрэглэх хэрэгтэй

• Өгөгдлийн нарийн төвөгтэй байдлаас шалтгаалан тусгай техник шаардлагатай болдог

• hexbin

• bigvis

BIGVIS

bigvis нь их хэмжээний өгөгөдлийн асуудлыг шийдвэрлэхийн тулд Hadley Wickham-ийн боловсруулсан туршилтын пакет байсан

• Hadley Wickham-ийн боловсруулсан Preprint болон R Meetup презентаци байдаг

• Бүтэн код нь дараахь линк дээр байгаа: https://github.com/hadley/bigvis-infovis

• Зорилт: 100 сая ажиглалтыг 5 секундад боловсруулах.

• Үндсэн зарчим: Дэлгэцний пикселээс илүү олон өгөгдлийн цэг шаардлагагүй.

• “ggstat” пакетыг эдгээр санааг нэгтгэх ирээдүйн төсөл гэж үзсээр ирсэн.

BIGVIS алхмууд

• Condense (bin, condense)

• Smooth (smooth, best_h, peel)

• Visualize (autoplot plus standard methods)

TRELLISCOPE (DELTARHO, WAS TESSERA)

Tessera-г Purdue, Pacific Northwest National Laboratory болон Mozilla хөгжүүлж ирсэн. 2014 оны 11 дүгээр сард гаргасан энэхүү төсөл нь олон тооны гэрээг байгуулаад байгаа.

• R-ийн орчинд ажиллуулснаар, Tessera кластер хооронд ажиллах, энэ орчинд дүн шинжилгээ хийх ачааллыг бууруулаж өгөх өөрийн командуудтай.

• datadr пакет нь хялбаршуулсан Hadoop-ийн интерфэйс бүхий MapReduce-тэй төстэй байдлаар “хуваах болон дахин хослуулах” үйлдлийг гүйцэтгэдэг.

• Tessera нь өөрийн гэсэн олон тооны хувьсагч, ажиглалтыг харуулж чадах Trelliscope визуалчлалын интерфэйстэй.

• Эхлэхийн тулд quickstart-ийг ажиллуулна

• Демо хувилбар нь энд байгаа.

Хэсэг 6: Дүгнэлт

Олж мэдсээр байх

Өгөгдлийн визуалчлал эцэс төсгөлгүй танин мэдэх үйл явцыг илэрхийлэ байна:

• програмчлалыг ашиглах

• өгөгдлийн гүн рүү нь орох

• интерактив байдлыг нэмэгдүүлэх

• ...хөгжилтэй байж, суралцсаар бай! [ж.нь., R-bloggers.com]

Өгөгдлийн визуалчлал => Ирээдүй

• Арга хэрэгсэл, пакетууд нь нарийн төвөгтэй, интерактив визуалчлалыг хялбарчилж өгдөг

• Стандарт, хүлээлтийг хөгжүүлэх нь өгөгдлийн визуалчлалыг хэм хэмжээ болгож өгөх болно

• Өгөгдлийн визуалчлалыг ойлгох нь өгөгдлийн/мэдээллийн literacy-ийн нэг бүрэлдэхүүн хэсэг

• Асуулт? Хэлэлцүүлэг?

Төгсгөл

• Асуулт байна уу?

• Маш их баярлалаа!

• Намайг дагаарай...

– https://youtube.com/librarianwomack

– https://www.linkedin.com/in/ryanwomack

– https://twitter.com/ryandata

– https://ryandata.wordpress.com